Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Ciência da Computação

Computação por Humanos na Perspectiva do Engajamento e Credibilidade de Seres Humanos e da Replicação de Tarefas

Lesandro Ponciano dos Santos

Tese submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Sistemas de Computação

Francisco Vilar Brasileiro (Orientador)

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

S237c Santos, Lesandro Ponciano dos.

Computação por humanos na perspectiva do engajamento e credibilidade de seres humanos e da replicação de tarefas / Lesandro Ponciano dos Santos. – Campina Grande, 2015.

163 f. : il. color.

Tese (Doutorado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2015.

"Orientação: Prof. Dr. Francisco Vilar Brasileiro".

Referências.

1. Computação por Humanos. 2. Engajamento. 3. Credibilidade. 4. Replicação. I. Brasileiro, Francisco Vilar. II. Título.

CDU 004.5(043)

"COMPUTAÇÃO POR HUMANOS NA PERSPECTIVA DO ENGAJAMENTO E CREDIBILIDADE DE SERES HUMANOS E DA REPLICAÇÃO DE TAREFAS"

LESANDRO PONCIANO DOS SANTOS

TESE APROVADA EM 23/11/2015

FRANCISCO VILAR BRASILEIRO, Ph.D, UFCG
Orientador(a)

HERMAN MARTINS GOMES, Ph.D, UFCG
Examinador(a)

NAZARENO FERREIRA DE ANDRADE, D.Sc, UFCG
Examinador(a)

JUSSARA MARQUES DE ALMEIDA, Dra., UFMG
Examinador(a)

RAFAEL DUARTE COELHO DOS SANTOS, D.Sc., INPE
Examinador(a)

CAMPINA GRANDE - PB

Resumo

Computação por humanos (*human computation*) é um modelo de computação que se baseia na coordenação de seres humanos para resolver problemas para os quais o sistema cognitivo humano é mais rápido ou preciso que os atuais sistemas computacionais baseados em processadores digitais. Em sistemas de computação por humanos, ao invés de máquinas, os processadores que realizam as computações são seres humanos. Usar adequadamente o poder cognitivo provido por tais seres humanos é fundamental para o sucesso desse tipo de sistema. Entretanto, pouco se sabe sobre as características de oferta de poder cognitivo e de como o sistema pode utilizar essa oferta de forma otimizada. Este estudo visa avançar esse conhecimento. Como referencial teórico-conceitual, propõe-se uma articulação de teorias e conceitos sobre computação por humanos, engajamento, credibilidade e otimização de desempenho. Considerando essa articulação, são propostas métricas para analisar a oferta de poder cognitivo em termos do engajamento e da credibilidade dos participantes. Como estudo de caso de estratégia de otimização de desempenho, propõe-se um algoritmo de replicação de tarefas que visa melhorar o uso do poder cognitivo levando em conta informações de credibilidade dos participantes. Por meio de análise de distribuições, correlações, regressões, classificação e agrupamento, os comportamentos de engajamento e credibilidade são caracterizados usando dados de seis sistemas reais. Entre os resultados obtidos, destacam-se diversos padrões comportamentais identificados na caracterização. Há duas classes de engajamento de participantes: os transientes, que atuam no sistema em apenas um dia e não retornam, e os regulares, que apresentam um engajamento mais duradouro. Os regulares são a minoria, mas são os mais importantes por agregarem maior tempo de computação ao sistema. Eles também não são homogêneos; subdividem-se em cinco grandes perfis, que podem ser rotulados como: empenhados, espasmódicos, persistentes, duradouros e moderados. A credibilidade dos participantes, por sua vez, pode ser medida usando várias métricas baseadas no nível de concordância entre eles. Tal credibilidade está negativamente correlacionada com a dificuldade das tarefas. Por fim, simulações do algoritmo de replicação proposto mostram que ele melhora o uso do poder cognitivo provido pelos participantes e permite tratar diversos compromissos entre diferentes requisitos de qualidade de serviço.

Abstract

Human computation is a computing approach that draws upon human cognitive abilities to solve computational tasks for which there are so far no satisfactory fully automated solutions. In human computation systems, the processors performing the computations are humans rather than machines. The effectiveness of this kind of system relies on its ability to optimize the use of the cognitive power provided by each human processor. However, little is known about how humans provide their cognitive power in these systems and how these systems can use such cognitive power properly. This study aims at advancing knowledge in this direction. To guide this study, we articulate a framework of theories and concepts about human computation, human engagement, human credibility, and the optimization of computational systems. Based on this theoretical-conceptual framework, we propose metrics to characterize the cognitive power available in a human computation system in terms of the engagement and the credibility of the participants. As case study of system optimization, we also propose a task replication algorithm that optimizes the use of the available cognitive power taking into account information about the credibility of participants. By using correlations, regressions, and clustering algorithms, we characterize the engagement and credibility of participants in data collected from six real systems. Several behavioral patterns are identified in such characterization. Participants can be divided into two broad classes of engagement: the transients, those who work in the system in just one day; and the regulars, those who exhibit a more lasting engagement. Regulars are the minority of participants, but they aggregate the larger amount of cognitive power to the system. They can be subdivided into five groups, labeled as: hardworking, spasmodic, persistent, lasting and moderate. The credibility of participants can be measured by using several different metrics based on the level of agreement among them. Regardless of the metric used, the credibility is negatively correlated with the degree of difficulty of the tasks. Results from simulation show that the proposed task replication algorithm can improve the ability of the system to properly use the cognitive power provided by participants. It also allows one to address trade-offs between different quality-of-service requirements.

Agradecimentos

Ao longo do desenvolvimento da pesquisa descrita neste documento, recebi sugestões e opiniões de diversos colegas no Laboratório de Sistemas Distribuídos, dentre os quais sou especialmente grato a Lívia Sampaio, Ianna Sodré, Nigini Abílio, Thiago Emmanuel, David Candeia, Marcus Carvalho e Raquel Lopes.

Muitos experimentos e análises reportados neste documento surgiram de conversas com a equipe que atua/atuou na construção e gerência da plataforma Contribua. Em particular, agradeço a Adabriand Furtado, Guilherme Gadelha, Jeymisson Oliveira, Ely Richardson, Mariana Souto e Jefferson Neves pelas inúmeras contribuições.

Muitas das análises apresentadas neste documento foram moldadas com sugestões recebidas de revisores de revistas e conferências aos quais resultados preliminares foram submetidos. Tais sugestões desempenharam um papel fundamental no resultado final que é apresentado neste documento. Em razão disso, sou muito grato aos revisores das revistas *Computing in Science and Engineering*, *Human Computation* e *Journal of Internet Services and Applications* e das conferências *XXXII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos* e *First AAAI Conference on Human Computation and Crowdsourcing*.

Sou grato a Jussara Almeida, Herman Martins, Nazareno Andrade e Rafael Santos pelos generosos comentários apresentados na versão preliminar deste documento e na ocasião da defesa da qualificação. Tais comentários foram de grande importância na busca de uma unidade, coesão e rigor tanto na condução da pesquisa como na elaboração desta versão do documento.

Agradeço a Francisco Brasileiro (Fubica) pela confiança de que superaríamos os inúmeros desafios que Computação por Humanos nos impõe e pela orientação na superação desses desafios.

A pesquisa reportada neste documento recebeu apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) nos anos de 2012, 2013 e 2014 por meio do Programa de Demanda Social. Em 2015, o apoio financeiro foi recebido do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) por meio de bolsa de Fixação de Recursos Humanos.

Conteúdo

Lista de Siglas

API	Acrônimo da expressão em inglês *Application Programming Interface*, geralmente traduzida como Interface de Programação de Aplicações.
BOINC	Acrônimo da expressão em inglês *Berkeley Open Infrastructure for Network Computing*. Trata-se de um sistema de *middleware* para sistemas de computação voluntária.
Bossa	Acrônimo da expressão em inglês *BOINC Open System for Skill Aggregation*. Trata-se de um sistema de *middleware* para sistemas de computação por humanos inspirado no sistema de *middleware* BOINC.
CERN	Acrônimo da expressão em francês *Conseil Européen pour la Recherche Nucléaire*, traduzida como Organização Européia para a Pesquisa Nuclear. Trata-se de uma organização européia que é dedicada a pesquisa nuclear.
FDA	Função de Distribuição Acumulada.
HIT	Acrônimo da expressão em inglês *Human Intelligence Task*, traduzida como Tarefas que Requerem Inteligência Humana. É sinônimo de tarefas de computação por humanos.
Mturk	Acrônimo de *Mechanical Turk*. Trata-se de um sistema de computação por humanos de propriedade da empresa Amazon.com, Inc.
NASA	Acrônimo da expressão em inglês *National Aeronautics and Space Administration*, geralmente traduzida como Administração Nacional da Aeronáutica e do Espaço. Trata-se de um organização do governo dos Estados Unidos.
QoS	Acrônimo da expressão em inglês *Quality of Service*, geralmente traduzido como Qualidade de Serviço.

reCAPTCHA	É um serviço utilizado para proteger sítios Web de ataques de *spammers* automatizados. O serviço é baseado no conceito de computação por humanos.

Lista de Símbolos

T	Conjunto de tarefas de uma aplicação.
t	Tarefa, $t \in T$. Cada tarefa possui um conjunto de itens de entrada e um conjunto de instruções sobre o trabalho a ser executado.
W	Conjunto de trabalhadores.
w	Trabalhador, $w \in W$.
j_w	Tempo decorrido entre o dia em que o trabalhador w se juntou ao projeto e o dia em que o projeto foi concluído.
A_w	Sequência de dias em que o trabalhador w esteve ativo executando tarefas.
D_w	Multiconjunto da quantidade de tempo que o trabalhador w dedicou ao sistema em cada dia em que esteve ativo.
B_w	Multiconjunto do número de dias decorridos entre cada dois dias sequenciais em que o trabalhador w esteve ativo.
a_w	Taxa de atividade do trabalhador w.
d_w	Quantidade de tempo dedicado diariamente pelo trabalhador w.
v_w	Variação na periodicidade do trabalhador w.
r_w	Duração relativa da atividade do trabalhador w.
h	Grau de dificuldade de uma tarefa.
H_w	Multiconjunto que contém os valores de dificuldades das tarefas (h) executadas pelo trabalhador w .
$K_{w,h}$	Conjunto composto pelos trabalhadores com os quais o trabalhador w concordou quando ele proveu uma resposta igual à resposta da maioria.
$k_{w,h}$	Credibilidade média dos trabalhadores no conjunto $K_{w,h}$.
$M_{w,h}$	Conjunto composto pelos trabalhadores que proveram a resposta da maioria quando o trabalhador w não proveu uma resposta igual à resposta da maioria.

$m_{w,h}$	Credibilidade média dos trabalhadores no conjunto $M_{w,h}$.
$c_{w,h}$	Credibilidade superficial do trabalhador w em tarefas com grau de dificuldade h.
$e_{w,h}$	Credibilidade experimentada do trabalhador w em tarefas com grau de dificuldade h.
$p_{w,h}$	Credibilidade presumida do trabalhador w em tarefas com grau de dificuldade h.
$r_{w,h}$	Credibilidade reputada do trabalhador w em tarefas com grau de dificuldade h.
$s_{w,t}$	Resposta gerada por um trabalhador w para uma tarefa t.
$\rho(x, y)$	Correlação de Spearman entre os valores nas listas x e y.
$d(x, y)$	Distância absoluta média entre os valores nas listas x e y.
$\tau(x, y)$	Distância de Kendall entre os ranques x e y.

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Nas últimas décadas, diversos estudos se dedicaram ao aumento do desempenho de sistemas de computação baseados em computadores digitais[1]. Como resultado, muito se avançou na arquitetura desses sistemas e no projeto e gerência das aplicações que eles executam. Tais avanços permitiram o emprego desses sistemas na solução de problemas cada vez mais complexos. No entanto, apesar desses avanços, ainda existem problemas que esses sistemas não podem resolver de forma eficaz e/ou eficiente (SAVAGE, 2012; BERNSTEIN; KLEIN; MALONE, 2012). São problemas para os quais os algoritmos atuais possuem ordem de complexidade proibitiva ou cujos algoritmos existentes são apenas heurísticas que geram soluções que apresentam acurácia abaixo da desejada pelos usuários do sistema.

Nos últimos anos, alguns estudos têm identificado que muitos desses problemas se adéquam ao sistema cognitivo humano de modo que podem ser resolvidos por seres humanos com maior velocidade e/ou precisão que os atuais sistemas computacionais baseados em computadores digitais (BERNSTEIN; KLEIN; MALONE, 2012; SAVAGE, 2012). Exemplos desses problemas são aqueles que surgem em domínios como processamento de linguagem natural (SNOW et al., 2008; BERNSTEIN et al., 2010; HU; BEDERSON; RESNIK, 2010), compreensão de conteúdo em imagens (AHN, 2005; BRANSON et al., 2010) e criatividade (YUEN; KING; LEUNG, 2011; ARAÚJO, 2013). Visando tirar proveito da capacidade cognitiva de seres humanos para resolver problemas nesses domínios, tem emergido um novo modelo de

[1]Neste documento o termo *computador digital* é empregado para designar toda máquina eletrônica projetada para realizar computações. Esse termo foi definido por Alan Turing em 1950. Ele permite distinguir computadores que são seres humanos dos computadores que são máquinas projetadas para realizar computação (TURING, 1950).

computação denominado *computação por humanos*[2].

Tarefas de computação por humanos seguem o conceito geral de computação, que consiste no mapeamento de entrada em saída por meio do processamento de um conjunto finito de instruções (TURING, 1950; LAW, 2011). Uma tarefa de computação por humanos consiste tipicamente de um conjunto de dados de entrada e um conjunto de instruções. A saída para a tarefa é gerada por um ser humano ao processar as instruções sobre os dados recebidos como entrada. Considere, por exemplo, uma tarefa projetada para identificar conteúdo adulto em imagens postadas por usuários em um sistema de compartilhamento de imagens. Nessa tarefa, o dado de entrada pode ser uma imagem e a instrução pode ser a seguinte mensagem: *Conteúdo adulto é todo aquele considerado impróprio para menores de 18 anos. A imagem abaixo apresenta esse tipo de conteúdo?* associada com as opções de resposta *"Sim"* e *"Não"*. Ao executar esse tipo de tarefa, o ser humano deve observar o dado de entrada, processar as instruções e prover uma resposta. A resposta provida é a saída da tarefa.

Diversos tipos de tarefas têm sido implementadas com base neste modelo de computação. Exemplos dessas tarefas são anotação do conteúdo de imagens (AHN; DABBISH, 2004), classificação do formato de galáxias a partir de imagens capturadas por telescópios (FORTSON et al., 2012), transcrição de informação textual contida em imagens digitalizadas de livros antigos (AHN, 2005; AHN et al., 2008), geração de conteúdo criativo sobre um tema determinado (ARAÚJO, 2013), etc. Os casos de sucesso no uso de soluções baseadas nesse modelo de computação têm motivado novas iniciativas de emprego de computação por humanos nos mais diversos contextos. Com isso tem surgido uma demanda de computação por humanos em larga escala.

Visando suprir essa crescente demanda, sistemas computacionais dedicados à execução de tarefas de computação por humanos têm sido desenvolvidos. Um sistema de computação por humanos pode ser modelado como um sistema computacional distribuído que orquestra o poder cognitivo de seres humanos, chamados de *trabalhadores*. Tal sistema agrega uma multidão de trabalhadores conectados à Internet e gerencia o poder cognitivo provido por eles de forma a executar tarefas que requerem inteligência humana. Dois tipos de sistemas

[2]É importante ressaltar que seres humanos já atuavam realizado computações antes mesmo do surgimento dos computadores digitais. Isso remonta ao período compreendido entre o início do século XVI e o meio do século XX. Uma breve descrição da computação antes dos computadores digitais é apresentada no Apêndice A e uma contextualização da expressão "computação por humanos" é apresentada no Apêndice B.

de computação por humanos bastante difundidos atualmente são os sistemas de *trabalho online* (IPEIROTIS; PROVOST; WANG, 2010) e os sistemas de *pensamento voluntário* (SANCHEZ et al., 2011).

Sistemas de trabalho *online* agregam trabalhadores que possuem uma motivação financeira. Dessa forma, nesse tipo de sistema, os trabalhadores ofertam o poder cognitivo em troca de uma remuneração. Um dos principais exemplos é a plataforma Amazon Mechanical Turk (Mturk[3]). Essa plataforma recebe diariamente[4] entre 40.000 e 50.000 tarefas para serem executadas e possui mais de 400.000 trabalhadores registrados (IPEIROTIS, 2010; ROSS et al., 2010; PONCIANO; BRASILEIRO, 2013). Sistemas de pensamento voluntário, por sua vez, agregam trabalhadores que executam tarefas sem esperar qualquer remuneração em troca. Um dos principais exemplos é a plataforma Zooniverse[5]. Essa plataforma agrega aproximadamente 1 milhão de trabalhadores cadastrados (SIMPSON; PAGE; ROURE, 2014) e já permitiu a execução de dezenas de milhões de tarefas (SAUERMANN; FRANZONI, 2015).

1.1 Problema

O sucesso de sistemas de computação por humanos está diretamente relacionado à sua habilidade de gerenciar adequadamente o poder cognitivo provido pelos trabalhadores. No entanto, pouco se sabe sobre as características de oferta de poder cognitivo e de como o sistema pode utilizá-lo de forma otimizada. Neste trabalho, aborda-se a oferta de poder cognitivo no sistema na perspectiva do engajamento cognitivo e da credibilidade dos trabalhadores.

Engajamento cognitivo (O'BRIEN; TOMS, 2008; SIMPSON, 2009; ATTFIELD et al., 2011; LEHMANN et al., 2012) e credibilidade (FOGG; TSENG, 1999; WATHEN; BUREL, 2002; RIEH; DANIELSON, 2007) têm sido estudados em diversos contextos e áreas do conhecimento. Entretanto, são aspectos ainda pouco tratados no contexto dos sistemas de computação por humanos. Nesses sistemas, engajamento cognitivo se refere ao comportamento do trabalhador em termos da duração do período de tempo em que ele permanece no sistema executando tarefas e da quantidade de tarefas executadas ao longo desse período. Credibilidade, por

[3]Página Web www.mturk.com. Último acesso em 01 de setembro de 2015.

[4]As submissões de tarefas no Mturk podem ser acompanhadas em tempo real no sistema Mturk Tracker http://www.mturk-tracker.com/#/arrivals, último acesso em 24 de setembro de 2015.

[5]Página Web www.zooniverse.org. Último acesso em 01 de setembro de 2015.

sua vez, refere-se a quão críveis são as respostas geradas pelos trabalhadores para as tarefas executadas por eles.

Os estudos sobre o engajamento cognitivo de trabalhadores em computação por humanos têm focado principalmente em analisar qualitativamente os fatores psicológicos de engajamento (RADDICK et al., 2008, 2010; ROTMAN et al., 2012; EVELEIGH et al., 2014). Esses estudos permitem entender características qualitativas do comportamento dos trabalhadores. Entretanto, infelizmente, eles não permitem analisar como se compõe a oferta de poder cognitivo no sistema, as características de atuação e a importância de cada trabalhador. Uma análise mais quantitativa pode permitir a identificação de perfis de trabalhadores por suas características de engajamento e a importância de cada perfil para o desempenho do sistema.

A credibilidade dos trabalhadores também tem sido pouco estudada nesses sistemas. O que existe na literatura são análises de fatores que podem levar os trabalhadores a não executarem as tarefas de forma satisfatória, como a remuneração oferecida (QUINN; BEDERSON, 2011; KITTUR et al., 2013) e a forma como as tarefas são projetadas (KOCHHAR; MAZZOCCHI; PARITOSH, 2010; EICKHOFF; VRIES, 2011; KAZAI; KAMPS; MILIC-FRAYLING, 2013). No entanto, pouco se sabe sobre como medir adequadamente a credibilidade de um trabalhador levanto em conta, por exemplo, seu histórico de atuação e a dificuldade das tarefas. Assim, há uma carência de métricas que sirvam de indicadores de credibilidade dos trabalhadores e que deem suporte à análise das respostas geradas por eles.

Na ausência de conhecimento sobre a credibilidade dos trabalhadores, a abordagem mais adotada para tratar incertezas nas respostas providas por eles é a obtenção e agregação de respostas redundantes (HOVY et al., 2013; SHESHADRI; LEASE, 2013; PONCIANO et al., 2014a). Nessa abordagem, cada tarefa é replicada para diversos trabalhadores. Após as respostas redundantes serem obtidas dos trabalhadores, aplica-se um algoritmo de agregação com tratamento de incertezas e se obtém a resposta final para a tarefa. O algoritmo mais simples e mais utilizado na prática é o voto majoritário, em que se considera correta a resposta mais frequente no conjunto de respostas redundantes (SHESHADRI; LEASE, 2013). A abordagem de replicação seguida de agregação gera diversos problemas. Um dos principais problemas é a definição da quantidade de respostas redundantes que terão que ser obtidas para cada tarefa. Se ela for subestimada, compromete-se a acurácia da resposta obtida na agregação. Por outro lado, se ela for superestimada, há um excesso de trabalho redundante e consequente

desperdício de poder cognitivo provido pelos trabalhadores.

Neste contexto, o **problema** tratado neste trabalho é o pouco conhecimento sobre as características de engajamento e credibilidade dos trabalhadores ao proverem poder cognitivo em sistemas de computação por humanos e de como esse conhecimento pode ser útil na otimização do desempenho desses sistemas. Como estudo de caso de tal otimização, investiga-se o uso da informação de credibilidade dos trabalhadores em algoritmos de replicação de tarefas. Ao tratar esse problema, este trabalho visa elucidar as características da oferta de poder cognitivo em sistemas de computação por humanos e melhorar a habilidade desses sistemas de usar adequadamente o poder cognitivo provido pelos trabalhadores.

1.2 Objetivos

O **principal objetivo** deste trabalho é investigar a tese de que um entendimento maior do engajamento e credibilidade em sistemas de computação por humanos permite caracterizar a oferta de poder cognitivo pelos trabalhadores e implementar estratégias que melhorem o uso do poder cognitivo provido por eles. Para atingir esse objetivo, os seguintes **objetivos específicos** devem ser alcançados:

1. Analisar as literaturas de engajamento de seres humanos, credibilidade de seres humanos e replicação de tarefas a fim de identificar os aspectos que devem ser considerados no contexto de sistemas de computação por humanos.

2. Delinear as principais características de engajamento dos trabalhadores ao proverem seu poder cognitivo. Trata-se de derivar métricas de grau e duração do engajamento cognitivo de trabalhadores em sistemas de computação por humanos.

3. Delinear as principais características de credibilidade dos trabalhadores ao proverem poder cognitivo no sistema. Trata-se de derivar métricas para medir a probabilidade de um trabalhador prover uma resposta crível para as tarefas que ele executa no sistema.

4. Propor um algoritmo de replicação de tarefas que faça uso de informações sobre credibilidade para tratar incertezas durante a execução das tarefas. O propósito é definir de forma adaptativa e otimizada o nível de redundância que precisa ser utilizado em cada tarefa.

1.3 Resultados e Contribuições

Em linhas gerais, os principais resultados e contribuições obtidos na pesquisa descrita neste documento são as seguintes:

- Propõe-se uma articulação teórica e conceitual sobre computação por humanos, engajamento de seres humanos, credibilidade de seres humanos e replicação de tarefas em uma perspectiva de desempenho de sistemas de computação por humanos. Mostra-se que, na perspectiva de um sistema distribuído, há três grandes dimensões a serem consideradas em computação por humanos, que são: os requisitos de qualidade de serviço dos usuários que submetem tarefas de computação por humanos para serem executadas; os aspectos humanos dos trabalhadores que executam as tarefas; e as estratégias de projeto e gerência de tarefas implementadas no sistema. Conceitua-se engajamento de seres humanos, credibilidade de seres humanos e replicação de tarefas no contexto de sistemas de computação por humanos e discute-se o papel que eles podem exercer no desempenho desse tipo de sistema. Além de atender a diversos propósitos, essa articulação constitui a lente conceitual e teórica por meio da qual se orienta a pesquisa descrita neste documento.

- Considerando a literatura sobre engajamento de seres humanos, são propostas métricas para medir o engajamento de trabalhadores em sistemas de computação por humanos. São propostas 4 métricas, definidas como: taxa de atividade, duração relativa da atividade, tempo dedicado diariamente e variação na periodicidade. Tais métricas se ajustam ao objetivo de analisar o grau de engajamento e a duração do engajamento dos trabalhadores. A taxa de atividade permite analisar a taxa de retorno de cada trabalhador ao sistema durante o período em que ele permanece atuando no sistema. O tempo dedicado diariamente dá uma visão da extensão do engajamento diário, que está relacionado com o período de duração do engajamento de curto prazo. A duração relativa da atividade permite analisar a duração do engajamento de longo prazo, ponderado pelo total de tempo em que o sistema permaneceu sendo observado. Finalmente, a variação na periodicidade informa como a periodicidade dos retornos ocorre.

- Considerando a literatura sobre credibilidade de seres humanos, são propostas mé-

tricas para medir a credibilidade dos trabalhadores. São 4 métricas definidas como: concordância experimentada, concordância reputada, concordância ponderada e concordância superficial. Essas métricas são definidas de modo a levar em consideração diferentes formas de medir a credibilidade e colocando maior ênfase em um ou outro aspecto da execução das tarefas. Concordância superficial é a métrica de credibilidade mais simples; ela leva em conta apenas o grau de concordância entres os trabalhadores. Concordância experimentada, por sua vez, mede o grau de concordância real ao levar em conta a concordância que seria esperada e deduzindo-se a quantidade de concordância que pode ocorrer devido ao acaso. Concordância presumida, por sua vez, pondera o grau de concordância com a quantidade de dados utilizados no seu cálculo. Finalmente, a métrica de credibilidade concordância reputada leva em conta não apenas a quantidade de concordância exibida por um trabalhador no passado, mas também a credibilidade dos outros trabalhadores com os quais ele concordou e discordou.

- Propõe-se um algoritmo de replicação de tarefas que faz uso de informações comportamentais dos trabalhadores e de estimativas da dificuldade das tarefas para otimizar métricas de qualidade de serviço de interesse do usuário do sistema, como: economia de réplicas e acurácia das respostas. Pode-se parametrizar requisitos de interesse como urgência, métrica de credibilidade a ser utilizada e nível de credibilidade requerida nas respostas. Estratégias de otimização multiobjetivo são consideradas para identificar configurações do algoritmo que apresentam melhor desempenho ao tratar o compromisso entre diferentes objetivos, tais como: economia de réplicas e acurácia das respostas.

A avaliação das métricas de engajamento e de credibilidade é realizada em estudos de caso que utilizam dados de seis projetos de computação por humanos bastante distintos: Galaxy Zoo, The Milky Way Project, Cell Spotting, Sun4All, Análise de Sentimentos e Julgamento de Fatos. Tal caracterização consiste na análise de semelhanças e diferenças entre os trabalhadores e de relações entre métricas por meio de análises de classificação, agrupamento, correlações e regressões. A avaliação do algoritmo de replicação, por sua vez, é realizada por simulações projetadas de forma a medir a capacidade do algoritmo proposto de otimizar o número de réplicas utilizadas e a acurácia das respostas considerando a vari-

ação de diversos parâmetros de interesse como nível de credibilidade requerida na resposta final, urgência de se obter uma resposta e métrica de credibilidade considerada. Utiliza-se otimização multiobjetivo na análise das melhores configurações do algoritmo.

Os resultados obtidos na caracterização revelam diversas características da oferta de poder cognitivo em computação por humanos que até então permaneciam desconhecidas. Usando as métricas de engajamento, observou-se que existem duas grandes classes de trabalhadores: transientes e regulares. Os trabalhadores transientes constituem uma maioria pouca engajada. Os trabalhadores regulares são minoritários, mas eles apresentam maior contribuição em termos do total de tempo de computação disponibilizado ao sistema. Trabalhadores nessa classe exibem 5 perfis de engajamento cognitivo, que podem ser rotulados como: empenhados, espasmódicos, persistentes, duradouros e moderados. Tais perfis diferem entre si em termos tanto das características de engajamento como do poder computacional agregado ao sistema. Cada perfil se destaca de forma positiva ou negativa em relação aos demais perfis em alguma métrica de engajamento. No geral, todos os perfis são pouco engajados em pelo menos uma métrica.

Os resultados obtidos na caracterização da credibilidade dos trabalhadores permitiram analisar diferentes formas de se estimar o quanto se pode acreditar em cada resposta obtida de trabalhadores em sistemas de computação por humanos. A credibilidade deles pode ser medida usando diferentes métricas baseadas no nível de concordância entre eles. Naturalmente, o valor da credibilidade tende a variar com a métrica utilizada, dependendo se ela é mais conservadora, como a concordância reputada, ou menos conservadora, como a concordância simples. A ordem de credibilidade dos trabalhadores também muda dependendo da métrica de credibilidade utilizada. Observa-se também que a credibilidade de cada trabalhador varia com a dificuldade da tarefa que ele executa e que os trabalhadores tendem a apresentar maior variação de credibilidade entre eles em tarefas de dificuldade moderada e difícil.

A avaliação do algoritmo de replicação de tarefas mostrou que é possível usar melhor o poder cognitivo provido pelos trabalhadores em sistemas de computação por humanos. O algoritmo de replicação adaptativa proposto permite que isso seja feito ao se obter economia de réplicas e respostas com alta acurácia enquanto atende a requisitos de credibilidade requerida e de urgência definidos como parâmetros para o algoritmo. Naturalmente, o de-

sempenho do algoritmo é altamente afetado por esses parâmetros. Estratégias de otimização permitem identificar os melhores parâmetros para se obter a maximização da economia de réplicas e da acurácia. No geral, o algoritmo de replicação proposto apresenta desempenho superior à replicação com nível de replicação fixo e que usa voto majoritário para eleger a resposta final para cada tarefa, que é amplamente utilizado em computação por humanos. O desempenho do algoritmo proposto também se aproxima do desempenho de um oráculo que sabe se um trabalhador proverá uma resposta correta ou incorreta.

Dessa forma, a pesquisa a que se refere este documento articula um importante arcabouço de teorias e conceitos relevantes à análise de computação por humanos em uma perspectiva de sistema distribuído (PONCIANO et al., 2014), um conjunto de métricas para analisar o engajamento de trabalhadores no sistema (PONCIANO et al., 2014b; PONCIANO; BRASILEIRO, 2014) e um conjunto de métricas para analisar a credibilidade de trabalhadores no sistema e otimizar a replicação de tarefas (PONCIANO; BRASILEIRO; GADELHA, 2013; PONCIANO et al., 2014a). Esse arcabouço mostra-se importante para raciocinar sobre o desempenho de sistemas de computação por humanos. Ao se explorar as áreas de engajamento, credibilidade e replicação, que até então não tinham sido tratadas no contexto de computação por humanos, esse trabalho abre diversas novas perspectivas de pesquisa.

1.4 Estrutura do Documento

Os conceitos tratados em computação por humanos envolvem estudos em diversas disciplinas, como, por exemplo, Ciência da Computação, Ciências Administrativas e Psicologia. O Capítulo 2 tem como objetivo apresentar uma contextualização de tais estudos e destacar os principais conceitos em computação por humanos que são relevantes à pesquisa descrita neste documento. Isso é realizado pela discussão dos conceitos de tarefas, aplicações, sistemas e estratégias de otimização de desempenho em sistemas de computação por humanos.

Após essa análise conceitual, apresentam-se os estudos da caracterização da oferta de poder cognitivo pelos trabalhadores. Esses estudos são apresentados em três capítulos, como segue: o Capítulo 3 apresenta o estudo do engajamento de trabalhadores; o Capítulo 4 apresenta o estudo da credibilidade dos trabalhadores; e o Capítulo 5 apresenta o estudo da relação entre o engajamento e a credibilidade dos trabalhadores e a dificuldade de tarefas. Con-

forme cada caso, em cada um desses capítulos, apresenta-se o arcabouço teórico-conceitual utilizado, a abordagem proposta, os métodos de avaliação utilizados e os resultados obtidos.

Após os estudos de caracterização da oferta de poder cognitivo, tem-se o estudo da otimização do poder cognitivo disponível por meio da replicação de tarefas. Esse estudo é apresentado no Capítulo 6. Nesse capítulo, discute-se o referencial teórico sobre replicação de tarefas e, à luz desse referencial, propõe-se um algoritmo de replicação de tarefas de computação por humanos. Os métodos de avaliação do algoritmo e os resultados obtidos na avaliação também são apresentados nesse capítulo.

Naturalmente, os estudos de caracterização da oferta de poder cognitivo e de otimização dessa oferta por meio da replicação de tarefas possuem limitações. Tais limitações são discutidas no Capítulo 7. Em seguida, no Capítulo 8, apresenta-se a análise dos trabalhos relacionados. Nessa análise, dá-se ênfase aos trabalhos que tratam, mesmo que de forma tangencial, algum aspecto relevante ao estudo de credibilidade de seres humanos, engajamento de seres humanos e desempenho de sistemas de computação por humanos. Finalmente, no Capítulo 9 são destacadas as principais conclusões e contribuições da pesquisa descrita neste documento. Além disso, discutem-se novas perspectivas de pesquisa que se abrem a partir da pesquisa reportada neste documento.

Capítulo 2

Contextualização do Ecossistema de Computação por Humanos

Computação por humanos é uma área de pesquisa em grande atividade em diversas disciplinas, como inteligência artificial (LAW; AHN, 2011), visão computacional (CROUSER; CHANG, 2012), interação humano-computador (QUINN; BEDERSON, 2011) e trabalho cooperativo auxiliado por computador (KITTUR et al., 2013). Cada uma dessas disciplinas põe ênfase nos aspectos de computação por humanos que são mais relevantes nos seus respectivos contextos.

Na área de inteligência artificial, muito se tem discutido sobre algoritmos que combinam inteligência de seres humanos e poder computacional de máquinas (LAW; AHN, 2011). Pesquisadores na área de visão computacional têm focado no estudo das capacidades humanas que permitem a efetiva recuperação de informação visual e de como elas podem informar o desenvolvimento de novos algoritmos que podem ser executados por máquinas (CROUSER; CHANG, 2012). Na disciplina de interação humano-computador, por sua vez, umas das principais correntes de estudo foca em como projetar instruções aos seres humanos de modo a reduzir erros e incentivar a participação (QUINN; BEDERSON, 2011). Finalmente, estudos em trabalho cooperativo auxiliado por computador têm focado na análise da interação entre seres humanos ao executarem tarefas e na análise dos comportamentos cooperativos e colaborativos que emergem nessa interação (KITTUR et al., 2013).

Muitas iniciativas em computação por humanos são contextualizadas em outra área de pesquisa denominada *crowdsourcing*. De forma geral, *crowdsourcing* se refere ao ato de pegar um atividade que tradicionalmente é realizada por uma pessoa sozinha e atribuí-la

a um grande grupo de pessoas sob a forma de um convite aberto (HOWE, 2006; QUINN; BEDERSON, 2011). As atividades realizadas pelos seres humanos em *crowdsourcing* podem ser as mais diversas e não necessariamente se basearem em habilidades cognitivas. Portanto, nem toda atividade definida como *crowdsourcing* pode ser definida como computação por humanos.

Ao presente trabalho, que tem maior ênfase na oferta de poder computacional, é relevante uma perspectiva pragmática que contextualize os conceitos relativos a um sistema computacional baseado no modelo de computação por humanos. Na ausência de uma organização da área de computação por humanos sob essa perspectiva, propôs-se uma organização como parte da pesquisa reportada neste documento. Tal organização é apresentada no restante desta seção pela análise dos conceitos de tarefas de computação por humanos, aplicações de computação por humanos, sistemas de computação por humanos e de estratégias de projeto, gerência e desempenho de sistemas de computação por humanos.

2.1 Tarefas de Computação por Humanos

O termo "computação" pode ser definido como o processo de mapear entrada em saída usando um conjunto explícito e finito de instruções (TURING, 1950; LAW, 2011). A expressão *computação por humanos*[1] (AHN, 2005; QUINN; BEDERSON, 2009, 2011; BERNSTEIN; KLEIN; MALONE, 2012) é utilizada neste trabalho para descrever o modelo de computação em que o mapeamento de entrada em saída é realizado por um ser humano fazendo uso de sua capacidade cognitiva (LAW, 2011; QUINN; BEDERSON, 2011; LAW; AHN, 2011). Assim, uma *tarefa* de computação por humanos consiste de um conjunto de dados de entrada e um conjunto de instruções que aos serem executadas permitem que uma resposta adequada seja gerada.

Os *dados de entrada* de uma tarefa de computação por humanos podem ser os mais diversos, como: imagens, vídeos, texto, áudio. As *instruções* da tarefa, por sua vez, se referem ao que deve ser feito com os dados de entrada de modo que uma resposta adequada seja gerada. Tais instruções podem ser de diversos tipos, por exemplo: transcrição do conteúdo dos

[1] "Computação por humanos" é uma tradução da expressão em inglês *human computation*. Existem autores que traduzem essa expressão como "computação humana". No Apêndice B, o termo "computação por humanos" e outros termos alternativos e correlatos existentes na literatura são discutidos.

dados de entrada, classificação deles em classes pré-definidas, ranqueamento segundo algum critério definido nas instruções, etc. As instruções não precisam ser atômicas. Elas podem consistir em um fluxo de ações a serem desempenhas. Ao executar uma tarefa, o ser humano observa os dados de entrada, executa as instruções sobre esses dados e gera uma saída. A saída gerada por ele é a *resposta* para a tarefa.

Tarefas podem ser classificadas como factuais ou não factuais, dependendo do grau de subjetividade de suas instruções. Tarefas são ditas não factuais quando as instruções envolvem muitos aspectos de subjetividade, como opinião, sentimento e criatividade. Por exemplo, define-se como não factual uma tarefa que exibe duas imagens do pôr do sol e pede ao ser humano que escolha a imagem que retrata o pôr do sol mais bonito. As respostas geradas pelos seres humanos nesse tipo de tarefa não são avaliadas em termos de corretude, i.e., não se define que uma resposta está correta ou incorreta. Em tarefas ditas factuais, por sua vez, as instruções são mais precisas. Em razão disso, as respostas podem ser avaliadas em termos de corretude por um ser humano especialista. Por exemplo, define-se como factual uma tarefa que exibe uma fotografia de uma paisagem e pergunta ao ser humano se existe ou não uma árvore na paisagem retratada na fotografia. Em tarefas desse tipo, busca-se uma resposta correta.

Tarefas de computação por humanos também podem diferir entre si pela complexidade e dificuldade a que expõe o ser humano. Embora existam semelhanças entre os conceitos de dificuldade e complexidade, eles não são nem independentes, nem equivalentes (LIU; LI, 2012). A complexidade de uma tarefa pode ser definida como a quantidade de esforço cognitivo que ela demanda. Dessa forma, a complexidade é um atributo exclusivo da tarefa e não depende do ser humano que a executa. Dificuldade, por sua vez, é um atributo do par tarefa e ser humano (ou grupo de seres humanos) que está executando a tarefa. Trata-se, portanto, de algo mais relacionado aos seres humanos do que à tarefa em si, como, por exemplo, familiaridade com o problema, quantidade de conhecimento e experiências anteriores. Assim, enquanto a percepção de complexidade é comum a todos os seres humanos que executarem uma tarefa, a percepção de dificuldade pode variar de um ser humano para outro. Além disso, a complexidade de uma tarefa é imutável, enquanto que a dificuldade de uma tarefa percebida por um determinado ser humano pode diminuir, ou até mesmo aumentar, quando avaliada em diferentes instantes de tempo. Tipicamente, o desempenho humano

diminui quando aumenta a dificuldade percebida ou a complexidade da tarefa (CALLISTER; SUWARNO; SEALS, 1992).

Para definir o quão complexa uma tarefa é, pode-se desenvolver estratégias que analisam os processos de execução, como a pergunta a ser respondida e os itens de entrada a serem processados. Entretanto, projetar um algoritmo capaz de estimar a complexidade de uma tarefa baseado nos dados de entrada pode ser tão complexo quanto projetar um algoritmo capaz de resolver a tarefa automaticamente. Estimar a dificuldade, por sua vez, requer considerar uma diversidade de fatores. Por exemplo, considere uma tarefa projetada para detectar ironia em mensagens de texto. Ela pode consistir da pergunta "A frase abaixo apresenta uma ironia?", e as opções de resposta podem ser "Sim" ou "Não". Grande quantidade dessas tarefas pode ser gerada coletando-se automaticamente frases em comentários feitos em notícias em páginas Web. Neste caso, não se tem qualquer controle do conteúdo de cada frase que será avaliada em cada tarefa por cada ser humano. Algumas frases podem ser fáceis de serem avaliadas por alguns seres humanos e outras serem muito difíceis.

Dessa forma, estimar a dificuldade de tarefas executadas por seres humanos é algo desafiador (GREITZER, 2005; LIU; LI, 2012). Por exemplo, tarefas difíceis podem requerer mais tempo para serem concluídas, mas o simples fato de uma tarefa ter consumido mais tempo não significa que ela é mais difícil. Tarefas que demandam a execução de muitos passos de computação também podem demandar muito tempo para que todos os passos sejam executados, sem necessariamente que os passos exijam muito conhecimento ou experiência. Considerando que a dificuldade afeta a acurácia das respostas geradas pelos seres humanos para as tarefas, o fato de diversos seres humanos divergirem sobre uma resposta final para uma tarefa é um indicador de que esse grupo de seres humanos está experimentando dificuldade para executá-la (GREITZER, 2005; ASLAM; PAVLU, 2007; ARCANJO et al., 2014; AROYO; WELTY, 2014).

2.2 Aplicações de Computação por Humanos

Tarefas de computação por humanos podem ser estruturadas em aplicações. Uma aplicação de computação por humanos é um conjunto de tarefas cujas soluções, uma vez agregadas, resolvem um problema de computação por humanos. Aplicações podem ser classificadas

como *workflow* ou *projeto* de acordo com a dependência existente entre as tarefas.

Uma aplicação é classificada como um *workflow* quando as tarefas que a compõem apresentam dependências entre si. Um exemplo típico de *workflow* de computação por humanos é implementado pela ferramenta de correção de texto Soylent (BERNSTEIN et al., 2010). Existem três tipos de tarefas nessa aplicação: *encontrar*, *corrigir* e *verificar*. A tarefa do tipo *encontrar* recebe como entrada um texto e pede ao ser humano que marque partes nesse texto que podem conter erros. A tarefa do tipo *corrigir* recebe como entrada o texto marcado nas tarefas do tipo *encontrar*. Elas pedem que os seres humanos corrijam eventuais erros existentes nas partes marcadas. Cada tarefa do tipo *corrigir* é executada por mais de um ser humano. Quando as respostas dessas tarefas são recebidas, é gerada uma tarefa do tipo *verificar*. Esse tipo de tarefa recebe como entrada o conjunto de respostas obtidas nas tarefas do tipo *corrigir*. Ela pede que o ser humano eleja nesse conjunto a melhor correção. Ao final dessa tarefa, tem-se o texto corrigido.

Aplicações compostas por tarefas independentes e que diferem entre si apenas em termos dos dados de entrada são definidas como *projeto*[2]. Por exemplo, uma grande quantidade de tarefas de detecção de ironia em mensagens de texto pode ser gerada coletando-se automaticamente frases em comentários feitos em notícias em páginas Web. Todas as tarefas podem ser iguais em termos das instruções e das opções de resposta, diferindo apenas quanto à frase a ser avaliada por cada ser humano. O conjunto de todas as tarefas constitui uma aplicação do tipo projeto. Esse tipo de aplicação é muito comum no ambiente científico, onde um dos grandes desafios é realizar uma mesma análise sobre um conjunto muito grande de dados. Esse é o caso, por exemplo, da análise de imagens de galáxias no projeto Galaxy Zoo (LINTOTT et al., 2008).

2.3 Sistemas de Computação por Humanos

Um *sistema de computação por humanos* é um sistema distribuído que permite a interação entre os *usuários*, que são seres humanos que possuem tarefas de computação por humanos para serem executadas, e os *trabalhadores*, que são seres humanos que executam tarefas. Como exemplificado na Figura 2.1, cada usuário e cada trabalhador que atua em um sis-

[2]Esse tipo de aplicação também é conhecido como saco-de-tarefas, da expressão em inglês *bag-of-tasks*.

tema de computação por humanos o faz por meio de um computador digital conectado à Internet (QUINN; BEDERSON, 2011).

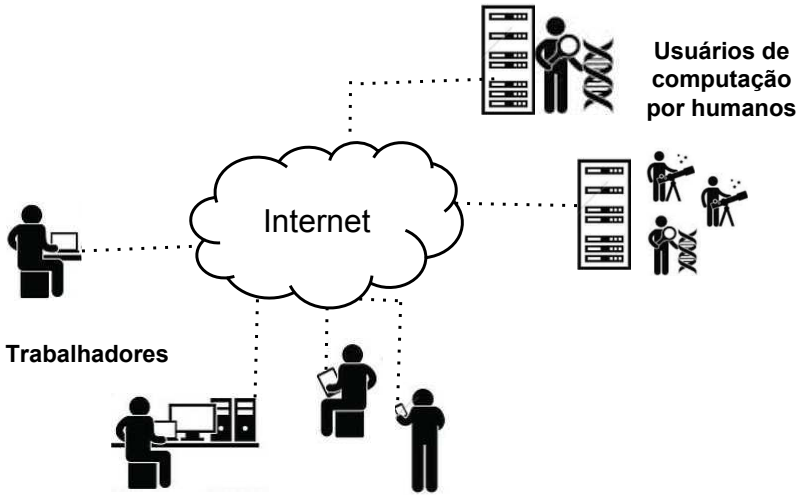


Figura 2.1: Ecossistema de computação por humanos. Destacam-se os usuários que submetem tarefas para serem executadas e os trabalhadores que executam as tarefas. Ambos atuam no sistema por meio de um computador digital conectado à Internet.

O sistema de computação por humanos implementa a interface entre os usuários e os trabalhadores. Esse tipo de sistema abstrai, para os usuários, grande parte da complexidade de fazer uso do poder cognitivo de uma multidão de trabalhadores. Para os trabalhadores, o sistema abstrai muito da complexidade de atuar em aplicações de diversos usuários. O sistema implementa diversas funcionalidades que visam gerenciar as tarefas providas pelos usuários e a capacidade cognitiva provida pelos trabalhadores conectados ao sistema. As funcionalidades implementadas tendem a variar com as características específicas de cada sistema.

Sistemas de computação por humanos podem ser projetados de modo que a execução de tarefas de computação por humanos não seja o foco principal do trabalhador no sistema. Dessa forma, a execução de tarefas de computação por humanos é apenas o subproduto decorrente de uma outra atividade. Isso ocorre, por exemplo, nos sistemas que implementam jogos com propósito (AHN; DABBISH, 2004) e no sistema reCAPTCHA (AHN et al., 2008). Jogos com propósito são projetados para que seres humanos contribuam com uma computação útil como um subproduto enquanto se entretêm jogando. reCAPTCHA[3], por sua vez, é um

[3]reCAPTCHA é um sistema de segurança. Ele se vale da premissa de que robôs não são capazes de trans-

sistema projetado para que seres humanos gerem uma computação útil enquanto validam o acesso à área restrita de um sítio Web.

Existem também sistemas de computação por humanos projetados de modo que o foco principal dos trabalhadores no sistema é executar tarefas de computação por humanos. Sistemas com essa característica podem ser subdivididos em mercados de trabalho *online* e sistemas de pensamento voluntário. Mercados de trabalho *online* agregam trabalhadores dispostos a executar tarefas de computação por humanos como um trabalho remunerado, enquanto sistemas de pensamento voluntário agregam trabalhadores dispostos a executar tarefas de computação por humanos como uma atividade voluntária (IPEIROTIS, 2010; YUEN; KING; LEUNG, 2011). Comparados aos sistemas em que o principal foco não é a execução de tarefas de computação por humanos, os mercados de trabalho *online* e os sistemas de pensamento voluntário são mais genéricos e versáteis em termos da diversidade de aplicações às quais podem dar suporte.

Em sistemas que implementam mercados de trabalho *online*, cada tarefa inclui um valor monetário a ser pago ao trabalhador quando ela é executada com sucesso. Dependendo do sistema, esse valor pode ser definido pelo usuário ao submeter as tarefas (como ocorre no sistema Mturk) ou definido em um leilão em que cada trabalhador define um preço a ser cobrado pela execução de uma tarefa (como ocorre no sistema UpWork[4]) (SATZGER et al., 2011). Cada tarefa também possui um tempo máximo em que ela pode permanecer alocada a um trabalhador, definido como duração da atribuição (*duration of an assignment*). Após esse tempo, caso não tenha executado a tarefa, o trabalhador perde o direito de executá-la e ela é disponibilizada novamente no quadro de tarefas disponíveis para seres executadas (*job board*). Quando uma tarefa é executada, o usuário que a submeteu pode aceitar a solução se a tarefa foi executada de forma satisfatória. Do contrário, ele pode rejeitá-la. Usualmente, quando um usuário rejeita a resposta para uma tarefa, ele compõe uma mensagem informando o problema na solução. Essa mensagem consiste em um *feedback* para o trabalhador que a executou. O trabalhador é remunerado apenas pelas tarefas cujas respostas

crever conteúdo existente em imagens e que seres humanos são. Assim, de modo a garantir que apenas seres humanos tenham acesso à área restrita de um sítio Web, o sistema exibe uma caixa de diálogo contendo uma imagem e só libera o acesso à área restrita se o conteúdo da imagem for transcrito corretamente. reCAPTCHA é projetado de forma que, ao transcreverem os conteúdos das imagens para terem acesso à área restrita, os seres humanos realizam uma computação útil como, por exemplo, transcrever o texto de livros antigos (AHN et al., 2008).

[4]Página Web www.upwork.com. Último acesso em 24 de setembro de 2015.

foram aceitas.

Sistemas de pensamento voluntário, por sua vez, agregam trabalhadores dispostos a executar tarefas sem receber em troca qualquer remuneração. Atualmente esse tipo de sistema tem sido desenvolvido principalmente no contexto de ciência cidadã[5]. Ciência cidadã consiste em uma parceria entre cientistas e pessoas sem especialização em atividades científicas, mas que estão dispostas a contribuírem na condução de uma pesquisa científica (SANCHEZ et al., 2011; LINTOTT; REED, 2013). Existe uma ampla diversidade de atividades na qual as pessoas podem contribuir em ciência cidadã (WIGGINS; CROWSTON, 2012). Uma dessas atividades é a execução de tarefas de computação por humanos. Em iniciativas de ciência cidadã que utilizam computação por humanos, os usuários são os cientistas que possuem tarefas de computação por humanos para serem executadas e os trabalhadores são as pessoas que desejam contribuir com a ciência executando tais tarefas.

Quando um usuário possui uma grande quantidade de tarefas de computação por humanos para serem executadas, é comum que ele crie um sistema de computação por humanos que é dedicado à execução das tarefas da sua aplicação. Em casos assim, como o sistema tem apenas uma única aplicação, o conceito de aplicação de computação por humanos e o conceito de sistema de computação por humanos se confundem. Nesses casos, aplicação e sistema são tratados de forma indistinta pelos trabalhadores. Isso é comum em aplicações de tipo projeto no qual tipicamente existem milhares e, em alguns casos, milhões de tarefas para serem executadas.

De todo esse contexto sobre computação por humanos, pode-se extrair que sistemas de computação por humanos são sistemas distribuídos heterogêneos e dinâmicos. A heterogeneidade pode se manifestar de diversas formas, por exemplo, nas diferenças entre os trabalhadores em termos de habilidades e dificuldades percebidas, assim como na diversidade de usuários e suas aplicações. O dinamismo, por sua vez, se manifesta na possibilidade de um trabalhador se juntar ao sistema ou se desligar dele em qualquer momento no tempo e nas variações de suas competências na execução das tarefas ao longo do período de atuação no sistema.

[5] A expressão ciência cidadã é uma tradução da expressão em inglês *citizen science.*

2.4 Desempenho em Sistemas de Computação por Humanos

Diversas estratégias têm sido propostas com o objetivo de se projetar aplicações, gerenciar componentes e otimizar o desempenho levando em consideração a heterogeneidade e o dinamismo de sistemas de computação por humanos. Estudos com esse propósito têm considerado as seguintes três dimensões: métricas de qualidade de serviço (QoS[6]), aspectos humanos dos trabalhadores e estratégias de projeto e gerência de aplicações. Cada uma dessas dimensões envolve uma diversidade de aspectos que podem ser considerados. Tais dimensões e os aspectos a elas associados são apresentas na Figura 2.2 e discutidos nos próximos parágrafos.

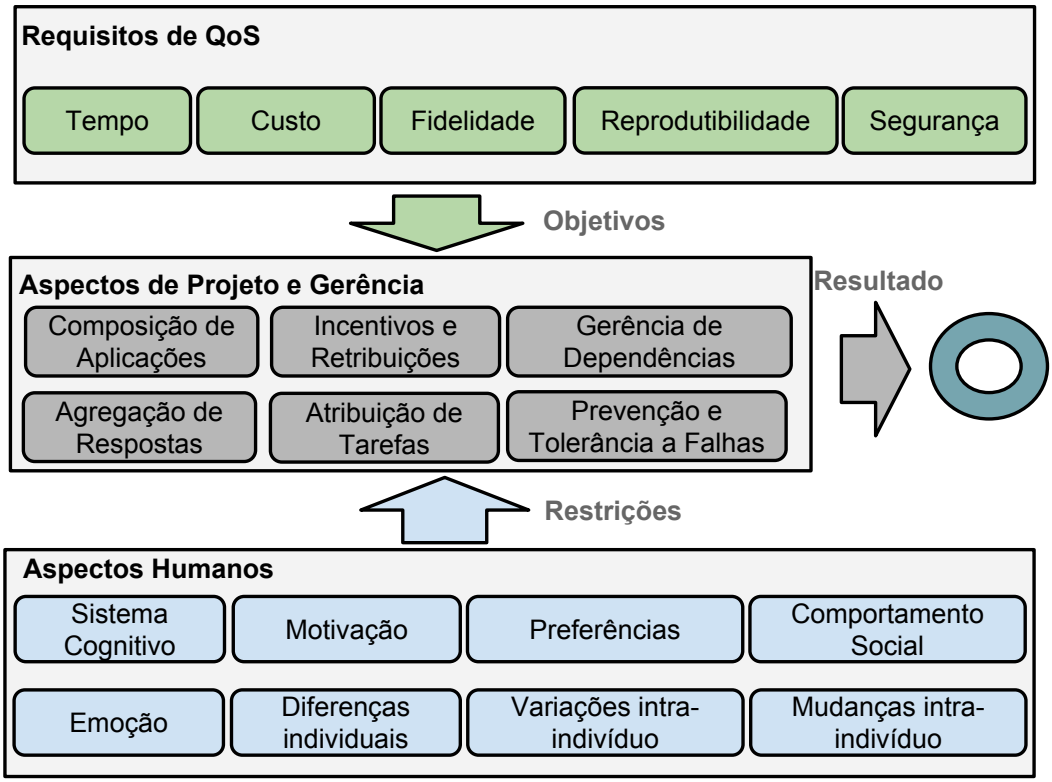


Figura 2.2: Aspectos considerados ao se avaliar e otimizar o desempenho de sistemas de computação por humanos. Destacam-se os requisitos de qualidade de serviço (QoS) dos usuários, os aspectos humanos dos trabalhadores e os aspectos de projeto e gerência de aplicações.

Ao submeterem aplicações para serem executadas, os usuários geralmente têm como

[6]Da expressão em inglês *quality of service*.

objetivo otimizar requisitos de qualidade de serviço. Tais requisitos estão geralmente relacionados ao tempo, custo, fidelidade, reprodutibilidade ou segurança.

- **Tempo.** Requisitos de tempo se referem à urgência na execução das tarefas. Eles incluem métricas como tempo de resposta, atrasos na execução acarretados por filas de tarefas, e limite máximo de tempo (*deadline*) em que a tarefa deve ser executada.

- **Custo.** Requisitos de custo se referem às despesas com a execução das tarefas. Ele é normalmente dividido em custos de promulgação e realização. Custo de promulgação são despesas com projeto das tarefas e custos de realização são despesas com a execução delas.

- **Fidelidade.** Requisitos de fidelidade estão associados ao quão bem uma tarefa é executada, considerando a correspondência entre as instruções definidas na tarefa e a resposta gerada pelo trabalhador (CARDOSO et al., 2002).

- **Reprodutibilidade.** O requisito de reprodutibilidade se refere a obter uma resposta semelhante para as tarefas se elas forem executadas novamente em diferentes momentos e/ou por diferentes grupos trabalhadores (PARITOSH, 2012).

- **Segurança.** Requisito de segurança diz respeito ao sigilo das tarefas e da confiabilidade dos trabalhadores.

Existem diversos aspectos humanos dos trabalhadores que precisam ser observados ao se otimizar o desempenho em sistemas de computação por humanos. Entre eles destacam-se: sistema cognitivo, motivação, preferências, comportamento social, emoção, diferenças individuais, variações intra-indivíduo e mudança intra-indivíduo.

- **Sistema cognitivo humano**. Sua função inclui vários processos de execução de tarefas pelos seres humanos, tais como processamento de informação, compreensão e aprendizagem. Ele especifica a organização de processos em memória de longo prazo e memória de curto prazo. A memória de longo prazo é o lugar onde o conhecimento é armazenado. Por sua vez, a memória de curto prazo é uma memória de trabalho usada para processar informação, no sentido de organizar, contrastar, e comparar (SIMON, 1990). Os seres humanos são capazes de lidar com alguns itens de informação

ao mesmo tempo em sua memória de curto prazo, e o processamento dos itens de informação também consomem espaço nessa memória (SWELLER; MERRIENBOER; PAAS, 1998). Quando o processamento dos itens de uma tarefa excede a capacidade de memória de curto prazo, têm-se o fenômeno denominado sobrecarga cognitiva.

- **Motivação**. Do ponto de vista da teoria geral da motivação (MASLOW, 1943), os seres humanos são guiados por impulsos ou objetivos, ou seja, o desejo de fazer/obter coisas novas e alcançar novas condições. Estudos de incentivos exploram a forma como esses objetivos influenciam o comportamento humano. Considerando a Teoria da Autodeterminação (DECI; RYAN, 2000), a motivação pode ser dividida em intrínseca e extrínseca. Na execução da tarefa, a motivação intrínseca consiste nos interesses intangíveis que levam os trabalhadores a executar uma tarefa específica, por exemplo, porque a tarefa lhe proporciona prazer ou lhe permite desenvolver uma habilidade particular. A motivação extrínseca, por sua vez, é composta por fatores externos aos trabalhadores, por exemplo, a quantidade de dinheiro que o trabalhador receberá por executar a tarefa.

- **Preferências**. Os seres humanos exibem preferências pessoais (KAPTEYN; WANSBEEK; BUYZE, 1978). Essas preferências são explicadas com base em dois tipos de influências: as suas próprias experiências passadas e as experiências de outros que são diretamente observáveis por eles. Como exemplo das preferências dos trabalhadores em tarefa, considere o caso em que, depois de se sentirem várias vezes entediados ao executarem tarefas que demandam muito tempo para serem concluídas, os trabalhadores passam a demonstrar preferência por executar apenas tarefas menos demoradas. Essa preferência pode ser percebida no sistema de computação por humanos, por exemplo, quando as tarefas menos demoradas são as primeiras a serem escolhidas.

- **Comportamento social**. Sociabilidade significa organização de grupos/comunidade para realizar atividades (COLEMAN, 1990). Em geral comunidades são formadas e persistem ao longo do tempo porque os indivíduos percebem vantagens em participar delas e, assim, elas servem aos seus interesses. A teoria de Senso de Comunidade sugere que os membros de uma comunidade desenvolvem senso de comunidade com base na participação, influência, integração e satisfação das suas necessidades. Eles também compartilham uma conexão emocional entre si (MCMILLAN, 1996). Em um

sistema de computação por humanos, esse comportamento pode influenciar a forma como os membros de uma comunidade se comportam ao executar as tarefas.

- **Emoção**. A emoção pode ser definida como um complexo estado psicológico e fisiológico que permite que os seres humanos detectem se um determinado evento é mais desejável ou menos desejável (DOLAN, 2002). Emoção se refere, por exemplo, ao humor, afeto, sentimento e opinião. Emoção interage com e influencia outros aspectos humanos relevantes para a eficácia de execução da tarefa. Por exemplo, a emoção influencia as funções do sistema cognitivo relacionadas à percepção, aprendizagem e raciocínio (DOLAN, 2002).

- **Variação intra-indivíduo**. Os seres humanos apresentam uma variabilidade no seu comportamento (RAM et al., 2005). Tal variabilidade é de curto prazo e não sistemática. Ela pode ser gerada por fatores como: oscilação, inconsistência e ruído.

- **Mudança intra-indivíduo**. Trata-se de uma mudança duradoura e sistemática no comportamento do indivíduo (RAM et al., 2005). Pode ser resultado, por exemplo, da aprendizagem e envelhecimento.

- **Diferenças individuais**. Os seres humanos apresentam uma variabilidade entre si em vários fatores (STANOVICH; WEST, 1998; PARASURAMAN; JIANG, 2012), como o processo de tomada de decisões e o desempenho apresentado nessas decisões. As diferenças individuais podem ser percebidas em três competências humanas: conhecimento, habilidades e proficiência. Conhecimento se refere a um corpo organizado de informações aplicado diretamente na execução de uma tarefa. Habilidade refere-se à capacidade de aplicar o conhecimento na execução das tarefas, normalmente medido qualitativa e quantitativamente. Proficiência são os comportamentos adequados na execução da tarefa propriamente dita e necessários para trazer conhecimento e habilidades na execução da tarefa.

Estratégias de projeto e gerência que focam em atingir requisitos de qualidade de serviço considerando os aspectos humanos dos trabalhadores podem ser mapeadas em seis grandes classes, são elas: composição de aplicações, gerência de dependências, atribuição de tarefas, agregação de respostas, tolerância a falhas e gerência de trabalhadores.

- **Composição de aplicações.** Estudos que focam nesse aspecto investigam estratégias para projetar tarefas de modo a tirar maior proveito do sistema cognitivo humano (KHANNA et al., 2010; CHANDLER; HORTON, 2011; KULKARNI; CAN; HARTMANN, 2012) e organizar as tarefas nas aplicações de modo a atingir determinadas métricas de desempenho (BERNSTEIN et al., 2010; LIN; MAUSAM; WELD, 2012; CHILTON et al., 2013; BRAGG; MAUSAM; WELD, 2013; BOZZON et al., 2013).

- **Incentivos e retribuições.** Envolve duas correntes de estudos. A primeira corrente tem ênfase em entender como definir os incentivos adequados para motivar os trabalhadores a executarem as tarefas, tais como incentivos financeiros, status e reconhecimento no sistema (MASON; WATTS, 2009; ARCHAK, 2010; ROGSTADIUS et al., 2011; SINGLA; KRAUSE, 2013; SINGER; MITTAL, 2013). A segunda corrente estuda esquemas de retribuição a serem utilizados para definir em que situações um trabalhador receberá a retribuição (WITKOWSKI et al., 2013; RAO; HUANG; FU, 2013). Um exemplo de esquema de retribuição é aquele em que o trabalhador é remunerado apenas se a resposta provida por ele for igual à resposta provida pela maioria dos trabalhadores que executarem a tarefa.

- **Gerência de dependências.** Esse aspecto se refere à execução eficiente de *workflows* de tarefas garantindo as dependências que podem existir entre elas. Estudos que focam nesse aspecto têm ênfase principal em: (i) delinear contextos em que se obtém ganho de desempenho ao permitir que trabalhadores que estão executando tarefas dependentes se comuniquem (MAO et al., 2011; ZHANG et al., 2012; IRANI; SILBERMAN, 2013) e (ii) identificar situações em que é melhor o trabalhador visualizar todo o estado do *workflow* e aquelas em que é mais adequado que ele tenha informação apenas da tarefa que ele está executando (MAO et al., 2011; KULKARNI; CAN; HARTMANN, 2012; KEARNS, 2012).

- **Atribuição de tarefas**. Estudos em atribuição de tarefas tratam da alocação de tarefas aos trabalhadores. De forma geral, existem duas abordagens: o sistema definir que tarefa cada trabalhador executará (MORRIS, 2011; NORONHA et al., 2011; SATZGER et al., 2011; SCHALL; SATZGER; PSAIER, 2014; DIFALLAH; DEMARTINI; CUDRÉ-MAUROUX, 2013) ou cada trabalhador escolher as tarefas que deseja executar (CHILTON et al., 2010;

AMBATI; VOGEL; CARBONELL, 2011). Em ambas as abordagens, diversas estratégias são possíveis. As estratégias geralmente consideram algum fator humano (como capacidade cognitiva e preferências) e/ou métricas de desempenho do sistema (como tempo de resposta e taxa de erros).

- **Agregação de respostas.** Esse aspecto trata da análise de um conjunto de respostas redundantes obtidas no sistema. Essa é uma atividade realizada após a aplicação terminar de executar (SHESHADRI; LEASE, 2013). De uma forma geral, o propósito da agregação varia com o tipo de tarefa. Em tarefas não-factuais, as estratégias de agregação geralmente tentam identificar as preferências ou opiniões dos trabalhadores que executaram as tarefas (YI et al., 2013; DALVI et al., 2013). Em tarefas factuais, por sua vez, as estratégias de agregação buscam identificar qual a resposta correta para cada tarefa (SHENG; PROVOST; IPEIROTIS, 2008; WHITEHILL et al., 2009; LITTLE et al., 2010; BAROWY et al., 2012; HOVY et al., 2013; WANG et al., 2013).

- **Prevenção e tolerância a falhas.** São estudos que analisam falhas que ocorrem durante a execução de tarefas e discutem como elas podem ser prevenidas, detectadas e tratadas (KOCHHAR; MAZZOCCHI; PARITOSH, 2010; IPEIROTIS; PROVOST; WANG, 2010; AMIR et al., 2013). Estudos sobre esse aspecto têm focado principalmente na prevenção de falhas. O tratamento de eventuais falhas dos trabalhadores que geram respostas incorretas tem sido feito por meio de estratégias de agregação de respostas.

A organização da literatura de computação por humanos apresentada nesta seção auxilia na análise dos diversos aspectos relacionados ao desempenho de sistemas de computação por humanos. Coloca-se em perspectiva as três dimensões que representam diferentes perspectivas em que é possível abordar computação por humanos quando se deseja aumentar o desempenho: os requisitos de QoS, os aspectos humanos e as estratégias de projeto e gerência. Cada dimensão está intimamente ligada a um agente no ecossistema de computação por humanos: os requisitos de QoS são medidas de interesse dos usuários; as estratégias de QoS estão relacionados à forma como os sistemas gerenciam a execução das aplicações; e os aspectos humanos são características dos trabalhadores que atuam executando tarefas no sistema. Cada uma dessas dimensões é composta por um conjunto de fatores que devem ser considerados. Tais fatores encontram-se destacados na Figura 2.2.

Considerando as suas relações, é evidente que as dimensões não são independentes. A definição de estratégias de projeto e gerência de aplicações é afetada por ambos os requisitos de QoS dos usuários e os aspectos humanos dos trabalhadores. Os requisitos de QoS dos usuários norteiam a elaboração de estratégias de projeto e gerência de aplicações. Aspectos humanos dos trabalhadores, por sua vez, delimitam um espaço onde as estratégias de projeto e gerência de aplicações podem agir com o objetivo de otimizar os requisitos de QoS. Dessa forma, o resultado final de uma aplicação executada em um sistema de computação por humanos é influenciado pelas três dimensões.

2.5 Considerações Finais

Neste capítulo, apresentou-se o ecossistema de computação por humanos por meio da discussão dos conceitos de tarefas, aplicações, sistemas de computação por humanos e de estratégias que têm sido propostas para melhorar o desempenho desses sistemas. A organização desse ecossistema e as análises apresentadas são contribuições da pesquisa reportada neste documento (PONCIANO et al., 2014). Análises como a apresentada neste capítulo podem desempenhar diversos papéis em pesquisas científicas e desenvolvimento de sistemas computacionais (GRUDIN; POLTROCK, 2012), como, por exemplo: *(i)* servir como um guia de aspectos que devem ser considerados por pesquisadores e desenvolvedores, *(ii)* prover uma linguagem comum por meio da qual os pesquisadores podem se comunicar e raciocinar sobre a área em estudo, e *(iii)* justificar e motivar o desenvolvimento de novas estratégias.

Uma vez apresentado o ecossistema de computação por humanos, pode-se então melhor localizar a pesquisa reportada neste documento. A Figura 2.3 destaca os aspecto tratados nesta pesquisa dentro do arcabouço de estudos em computação por humanos descrito neste capítulo. Neste trabalho, dá-se ênfase a dois comportamentos humanos relacionados à oferta de poder cognitivo: engajamento e credibilidade. Esses dois comportamentos são caracterizados a fim de entender a oferta de poder cognitivo em projetos de computação por humanos. Foca-se em apenas aplicações do tipo projeto e dá-se ênfase a cada projeto como um sistema de computação por humanos. Como um estudo de caso da otimização do uso que se faz do poder cognitivo disponível, analisa-se o uso de informações de credibilidade na agregação de respostas e na tolerância a falhas por meio de estratégias de replicação de tarefas. Todo o tra-

balho é conduzido tendo em vista requisitos de qualidade de serviço associados às métricas de tempo, custo e fidelidade.

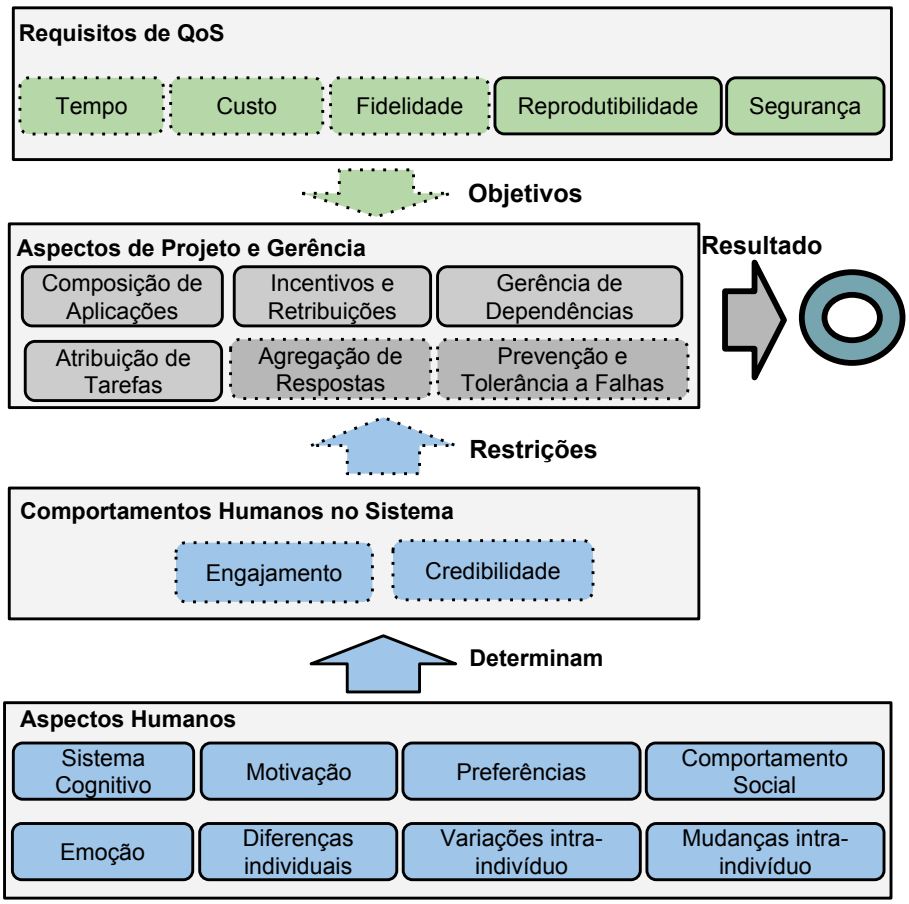


Figura 2.3: Engajamento e Credibilidade no universo de aspectos que têm sido tratados em computação por humanos. Os níveis enfatizados nesta pesquisa estão destacados com margens pontilhadas, são eles (i) os requisitos de QoS: tempo, custo e fidelidade; (ii) os aspectos de projeto e gerência: agregação de respostas e tolerância a falhas; e (iii) os comportamentos humanos: engajamento e credibilidade.

Engajamento de seres humanos, credibilidade de seres humanos e replicação de tarefas são grandes áreas que têm sido objetos de estudo em diversas disciplinas. Há uma ampla literatura de conceitos e teorias desenvolvidas nessas áreas que servem de base teórica para o estudo conduzido neste trabalho. Essa literatura é revisada nos próximos capítulos, quando se detalha os estudos realizados em engajamento, credibilidade e replicação de tarefas em projetos de computação por humanos.

Capítulo 3

Engajamento de Trabalhadores em Projetos de Computação por Humanos

Um dos requisitos fundamentais para o sucesso de um projeto de computação por humanos é existirem pessoas dispostas a se juntarem a ele e permanecerem executando tarefas ao longo de algum período de tempo. Em razão disso, entender o comportamento de tais pessoas ao proverem o poder cognitivo é fundamental quando se deseja entender e melhorar o desempenho desse tipo de projeto. Neste trabalho, defende-se que tal comportamento pode ser analisado à luz da literatura que trata do engajamento de seres humanos. Dessa forma, neste capítulo, descreve-se a pesquisa feita para alcançar um melhor entendimento da atuação de trabalhadores em projetos de computação por humanos usando como lente o conceito de engajamento de seres humanos.

A pesquisa parte da investigação de duas questões de pesquisa gerais sobre o que é engajamento no contexto de computação por humanos e como quantificar esse engajamento. A partir desse entendimento, busca-se delinear, usando dados de projetos reais, quais são as principais características de engajamento dos trabalhadores em projetos de computação por humanos e de que forma os trabalhadores diferem entre si em termos do padrão de engajamento e do poder cognitivo agregado ao projeto.

Nas seções seguintes, primeiro apresenta-se uma contextualização de engajamento (Seção 3.1). Após isso, propõe-se métricas para medir o engajamento dos trabalhadores (Seção 3.2). Finalmente, apresenta-se a avaliação realizada usando dados de projetos reais, detalhando-se os materiais e métodos de avaliação (Seção 3.3) e os resultados obtidos (Se-

ção 3.4).

3.1 Fundamentos do Engajamento de Seres Humanos

Busca-se nesta seção prover um melhor entendimento dos conceitos associados à literatura sobre engajamento de seres humanos, como a definição de engajamento, dos tipos de engajamento, formas de avaliação do engajamento e de fatores que determinam o engajamento de seres humanos.

3.1.1 O que é Engajamento?

O termo engajamento tem sido empregado em estudos em diversas disciplinas, como: Educação, Ciências Administrativas e Ciência da Computação. Em muitos estudos, ele é definido de forma diferente. Bakker e Demerouti (2008) estudam engajamento de seres humanos ao desempenhar atividades em organizações. Eles definem engajamento como um estado positivo e realizador, caracterizado pelo vigor, dedicação e absorção. Vigor consiste no alto nível de energia e resiliência durante o desempenho de uma atividade. Dedicação se refere a um forte envolvimento com a atividade e pela experiência de uma sensação de significância, entusiasmo e desafio. Finalmente, absorção ocorre quando se está totalmente concentrado em uma atividade. Rodden, Hutchinson e Fu (2010) e Lehmann et al. (2012), por sua vez, estudam engajamento em sistemas Web. Rodden, Hutchinson e Fu definem engajamento como o nível de envolvimento de um ser humano com um sistema, enquanto Lehmann et al. definem como a qualidade de experiência de um ser humano que enfatiza o aspecto positivo da interação e um fenômeno particular associado com estar cativado pelo sistema, estar motivado a usá-lo.

Dadas as diferentes formas como o termo engajamento tem sido definido, surgiram estudos com o propósito de discutir as diversas definições de engajamento e organizar um arcabouço conceitual que dê suporte à definição e que organize uma perspectiva de pesquisa interdisciplinar. O'Brien e Toms (2008) propõem um arcabouço conceitual para definir o engajamento de seres humanos com tecnologia. Eles definem engajamento como a qualidade de experiência que é influenciada por atributos como estética, apelo sensorial, motivação e interesse. Enquanto comportamento, o engajamento é modelado como um processo com-

posto de quatro estados: *ponto de engajamento*, que é o instante de tempo quando o ser humano se junta ao sistema; *período de engajamento contínuo*, que é o período contínuo de tempo em que o ser humano permanece interagindo com o sistema; *desengajamento*, que corre quando um período de engajamento é finalizado; e *reengajamento*, quando o ser humano inicia um novo ponto de engajamento com o sistema.

3.1.2 Tipos de Engajamento

O *tipo do engajamento* do ser humano em um sistema é definido pelo tipo de recurso que ele investe ao atuar no sistema. Exemplos de tipos de engajamento são o engajamento social e o engajamento cognitivo. O engajamento social ocorre quando a atuação do ser humano no sistema consiste em atividades que envolvem a interação com outros seres humanos. Pessoas que possuem dificuldade de interação social podem ter dificuldade de se engajar e permanecerem engajados ao longo do tempo nesse tipo de atividade. Engajamento cognitivo se refere às ações que exigem que o ser humano realize esforço cognitivo como operações matemáticas, seleção, agrupamento e ranqueamento de itens. Pessoas que não gostam de realizar esse tipo de atividade podem não apresentar um engajamento cognitivo sustentável.

Engajamento social é amplamente estudado em áreas como as redes sociais *online* e comunidades virtuais (MILLEN; PATTERSON, 2002). Em alguns sistemas de computação por humanos, trabalhadores podem realizar atividades que podem ser caracterizadas como engajamento social. Isso ocorre, por exemplo, quando eles interagem entre si em fóruns (FORTSON et al., 2012). No entanto, o principal tipo de engajamento dos trabalhadores em computação por humanos é o engajamento cognitivo no processo de execução de tarefas. Esse tipo de engajamento tem sido amplamente abordado em psicologia educacional e engajamento no trabalho (GONZÁLEZ-ROMÁ et al., 2006; SIMPSON, 2009).

A distinção de tipo de engajamento implica em levar em conta que o processo de engajamento do ser humano com o sistema pode diferir dependendo do que ele faz no sistema. Assim, os padrões observados em um processo de engajamento social podem ser bastante diferentes dos observados em um processo de engajamento cognitivo. Dada a ênfase no estudo de computação por humanos, este trabalho se concentra na análise do engajamento dos seres humanos em atividades cognitivas.

3.1.3 Avaliação do Engajamento

Engajamento pode ser medido por meio de medidas subjetivas ou objetivas (ATTFIELD et al., 2011). Medidas subjetivas são obtidas por meio da aplicação de questionários e/ou condução de entrevistas com os participantes do sistema. Medidas objetivas, por sua vez, podem ser obtidas, por exemplo, por meio de medição do tempo de permanência e taxa de retorno do participante ao sistema. Medidas objetivas são especialmente importantes quando se deseja construir um sistema capaz de detectar automaticamente o comportamento do participante e reagir de alguma forma a esse comportamento (FISCHER, 2001). Uma análise objetiva do engajamento distingue grau e duração do engajamento.

A *duração do engajamento* é uma medida de retenção do ser humano. Ela consiste basicamente em quanto tempo o ser humano permanece interagindo com o sistema. Pode ser amplamente classificada como engajamento de curto prazo ou engajamento de longo prazo. O engajamento é dito de curto prazo quando ele ocorre em um período de tempo relativamente curto, tal como minutos ou horas. Do contrário, ele é classificado como engajamento de longo prazo quando dura por um longo período de tempo, tal como meses ou anos. Em ambas as situações, seguindo o arcabouço proposto por O'Brien e Toms, o engajamento consiste de um ponto de engajamento, um período de engajamento sustentado, e um ponto de desengajamento.

No engajamento de curto prazo, o ponto de engajamento é o instante de tempo em que o humano realiza a primeira ação no sistema. O período de engajamento é o período de tempo durante o qual ele interage com o sistema continuamente. Finalmente, o ponto de desengajamento, ocorre quando o ser humano realiza a última ação finalizando o período de interação contínua. No engajamento de longo prazo, por sua vez, o ponto de engajamento é o instante de tempo em que o ser humano realiza a primeira ação no sistema. O período de engajamento consiste nos dias durante os quais ele continua interagindo com o sistema. Por fim, o ponto de desengajamento ocorre no dia em que ele deixa de usar o sistema definitivamente. Nessa perspectiva, o engajamento de longo prazo pode ser constituído por vários ciclos de engajamento de curto prazo.

O *grau de engajamento* é uma medida da participação dos seres humanos durante o período de engajamento. Também pode ser visto como uma medida da quantidade de recurso aplicado pelo ser humano ao interagir com o sistema. Medir o grau de engajamento tem se

mostrado uma tarefa desafiadora nos mais diversos tipos de sistemas (RODDEN; HUTCHINSON; FU, 2010; LEHMANN et al., 2012). Alguns estudos utilizam dados comportamentais armazenados em dados históricos do sistema para medir o grau de envolvimento dos seres humanos com o sistema. Estudos com base nessa abordagem geralmente consideram métricas, como: frequência de visitas, número de tarefas executadas, e quantidade de tempo interagindo com o sistema (RODDEN; HUTCHINSON; FU, 2010; LEHMANN et al., 2012).

3.1.4 Determinantes do Engajamento

Diversas teorias podem ser instanciadas para tentar explicar o engajamento de seres humanos em um sistema. No contexto do engajamento cognitivo, uma teoria que geralmente mostra-se relevante é a teoria da auto-eficácia (BANDURA, 1977). Essa teoria propõe explicações sobre as relações entre tempo e desempenho no comportamento de execução de tarefa por seres humanos. Ela afirma principalmente que a percepção humana de auto-eficácia determina se ele se engajará em uma atividade, o quanto de esforço será despendido e por quanto tempo a atividade será sustentada. Quanto mais um ser humano percebe que ele está desempenhando uma atividade de forma satisfatória, maior é a propensão de ele continuar. Assim, relaciona-se, por exemplo, o número de tarefas que um ser humano executa e percepção dele de que as tarefas foram executadas corretamente.

Naturalmente, os diversos aspectos humanos destacados na Figura 2.3 individualmente ou em conjunto podem exercer algum efeito sobre o grau e a duração do engajamento dos trabalhadores em sistemas de computação por humanos (PONCIANO et al., 2014). Identificar os aspectos que se provam relevantes em cada contexto é uma um objetivo bastante relevante e que tem guiado diversos estudos. Muitos desses estudos são discutidos ao longo deste documento. Entretanto, é importante ressaltar que a pesquisa conduzida neste documento tem maior ênfase em caracterizar o engajamento do que em isolar os fatores que os determinam.

3.2 Medindo Engajamento Cognitivo em Computação por Humanos

Neste trabalho, os trabalhadores são caracterizados considerando como eles se comportam em métricas de engajamento específicas. Métricas de engajamento são medidas de interação e envolvimento do trabalhador com o projeto. As métricas de engajamento propostas nesta seção são baseados no arcabouço conceitual proposto por O'Brien e Toms (2008). Ao utilizar este arcabouço, dá-se ênfase no engajamento dos trabalhadores ao longo do tempo, levando em conta os seus pontos de engajamento, períodos de contínuo engajamento, desengajamentos e reengajamentos.

A Figura 3.1 mostra uma visão geral do comportamento de um trabalhador ao longo do tempo. Esta figura mostra cinco conceitos utilizados nos cálculos das métricas: total de tempo que o trabalhador poderia permanecer ligado ao projeto, total de tempo que o trabalhador permaneceu ligado ao projeto, os dias ativos, o tempo dedicado em um dia ativo, e o tempo decorrido entre cada par de dias ativos. É conhecido que, em sistemas Web, grande parte dos seres humanos visita o sistema uma única vez e não retorna mais (ARGUELLO et al., 2006). Muitos deles visitam o sistema apenas por curiosidade e podem não ter interesse em um engajamento de longo prazo. Este trabalho distingue tais comportamentos em sistemas de computação por humanos e foca em uma análise mais aprofundada do engajamento dos trabalhadores que executam tarefas em pelo menos dois dias distintos. Dessa forma, as métricas são projetadas para medir o engajamento dos trabalhadores que apresentam uma atuação permanente, tendo atuado em pelo menos dois dias diferentes. Ao fazer isso, este trabalho dá maior ênfase à análise do engajamento de trabalhadores que tendem a apresentar interesse maior do que apenas uma curiosidade.

Um *dia ativo* de um trabalhador w é um dia em que este trabalhador esteve ativo executando tarefas no projeto. Considera-se que um trabalhador esteve ativo em um determinado dia, se ele executou pelo menos uma tarefa durante esse dia. Define-se A_w como um conjunto de datas em que o trabalhador esteve ativo, portanto, o número dias em que o trabalhador esteve ativo é dado por $|A_w|$. O *tempo que o trabalhador w permaneceu no projeto* é o número dias decorridos entre o primeiro dia ativo no projeto (i.e. $min(A_w)$) e o último dia ativo no projeto (i.e. $max(A_w)$). Formalmente, esse tempo é calculado como

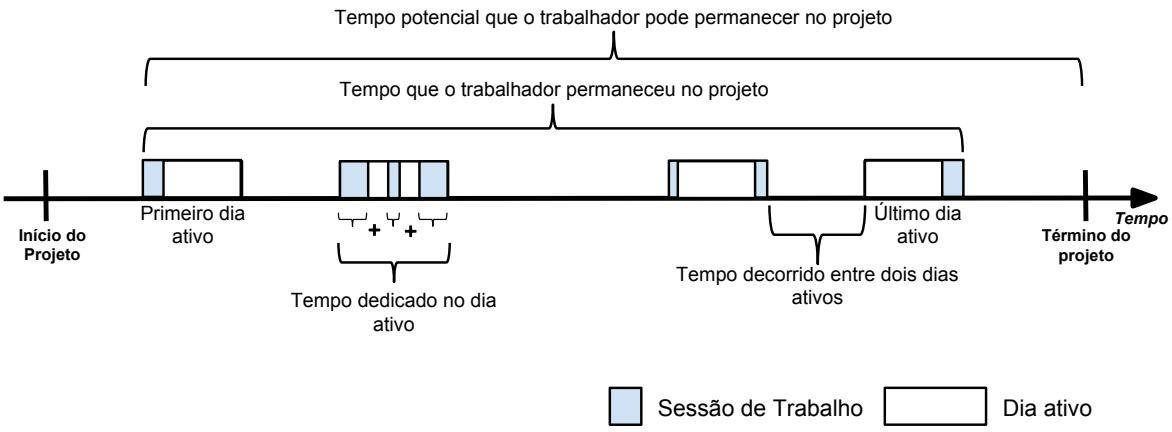


Figura 3.1: Linha de tempo da atuação de um trabalhador em um projeto. Destacam-se as informações utilizadas no cálculo das métricas de engajamento, tais como os dias ativos e as sessões de trabalho.

$max(A_w) - min(A_w) + 1$. O *total de tempo que um trabalhador* w *potencialmente pode permanecer ligado ao projeto*, por sua vez, é o número de dias decorridos entre o dia em que o trabalhador se juntou ao projeto (primeiro dia ativo) e o dia em que o projeto é concluído. Esse total de tempo é definido como j_w dias.

O *total de tempo dedicado em um dado dia ativo* é a soma da duração das sessões nesse dia. Sessões de contribuição são curtos períodos contínuos de tempo durante o qual o trabalhador permanece executando tarefas. Define-se D_w como o multiconjunto da quantidade de tempo que o trabalhador w dedica ao projeto em cada dia ativo. O *tempo decorrido entre dois dias ativos* é o número de dias que o trabalhador gastou para voltar ao projeto desde o último dia ativo. Define-se B_w como o multiconjunto do número de dias decorridos entre cada dois dias ativos sequenciais. Considerando j_w, A_w, D_w e B_w, pode-se derivar as métricas para medir o grau de engajamento (participação) e a duração do engajamento (retenção) de cada trabalhador.

São derivadas duas métricas de grau de engajamento: taxa de atividade e tempo dedicado diariamente. **Taxa de atividade (a_w)** é a proporção de dias em que o trabalhador esteve ativo em relação ao total de dias que ele permaneceu ligado ao sistema. Esse valor é calculado como indicado na Equação 3.1. Quanto mais próximo de 1, mais assíduo o trabalhador foi durante o tempo em que ele permaneceu ligado ao projeto.

$$a_w = \frac{|A_w|}{(max(A_w) - min(A_w)) + 1}, a_w \in (0, 1] \tag{3.1}$$

Tempo dedicado diariamente (d_w) é o total de tempo em horas que o trabalhador permaneceu em média executando tarefas em cada dia em que ele esteve ativo. Essa métrica é calculada como indicado na Equação 3.2. Quanto maior essa média, maior o tempo que o trabalhador dedica executando tarefas no sistema em cada dia ativo. Note que, porque tarefas de computação por humanos podem requerer diferentes quantidades de tempo de computação, a quantidade de tempo que os trabalhadores dedicam executando tarefas é uma medida de dedicação ao sistema melhor do que o número de tarefas que eles executam (GEIGER; HALFAKER, 2013; PONCIANO et al., 2014b).

$$d_w = avg(D_w), d_w \in (0, 24] \tag{3.2}$$

São definidas duas métricas para analisar a duração do engajamento: duração relativa da atividade e variação na periodicidade. **Duração relativa da atividade** (r_w) é a razão entre o total de dias em que o trabalhador permaneceu ligado ao projeto e o total de tempo que um trabalhador potencialmente pode permanecer ligado ao projeto (j_w). Dessa forma, essa métrica pode ser calculada como indicado na Equação 3.3. Quando $r_w = 1$, o trabalhador permanece no projeto durante todo o período decorrido entre o dia em que se juntou ao projeto e o dia em que o projeto foi concluído. Quanto mais próximo de 1, mais persistente o trabalhador foi no projeto.

$$r_w = \frac{max(A_w) - min(A_w) + 1}{j_w}, r_w \in (0, 1] \tag{3.3}$$

Variação na periodicidade (v_w), por sua vez, é o desvio padrão dos tempos decorridos entre cada par de dias ativos sequenciais. Ele é calculado como indicado na Equação 3.4. Quando $v_w = 0$, o trabalhador apresenta variação constante no tempo decorrido entre cada par de dias ativos sequenciais, o que indica que ele volta ao sistema com periodicidade perfeita ($sd = 0$). De modo inverso, quanto maior v_w, mais inconstante e menos previsível é a periodicidade com que o trabalhador volta ao sistema para executar mais tarefas.

$$v_w = sd(B_w) \tag{3.4}$$

As métricas de engajamento acima se ajustam ao objetivo de analisar o grau de engajamento e a duração do engajamento dos trabalhadores. A taxa de atividade permite analisar a taxa de retorno de cada trabalhador ao sistema durante o período em que ele permanece contribuindo. O tempo dedicado diariamente dá uma visão da extensão do engajamento diário, que está relacionada com o período de duração do engajamento de curto prazo. A duração relativa da atividade permite analisar a duração do engajamento de longo prazo, ponderado pelo total de tempo em que o sistema permaneceu sendo observado. Finalmente, a variação na periodicidade informa como a periodicidade dos retornos ocorre.

As quatro métricas propostas constituem diferentes dimensões do engajamento. A fim de identificar grupos de trabalhadores que são semelhantes entre si nessas quatro dimensões, utiliza-se algoritmos de agrupamento. A entrada para algoritmos de agrupamento é uma matriz $|W| \times 4$ em que cada linha representa um trabalhador $w \in W$ e cada coluna é uma métrica de engajamento, ou seja, a, d, r, e v. Como os resultados do agrupamento dependem dos valores relativos dos parâmetros a serem agrupados, uma normalização dos parâmetros antes do agrupamento é desejável (JAIN, 2008). Utiliza-se a normalização por intervalo para dimensionar os valores das métricas de engajamento no intervalo $[0, 1]$. A fórmula de normalização é definida como $x_w = \frac{x_w - x_{min}}{x_{max} - x_{min}}$, onde x denota a métrica de engajamento e w o trabalhador.

Para identificar o número adequado de grupos, primeiro executa-se um algoritmo de agrupamento hierárquico e observa-se o seu dendrograma, que produz um intervalo adequado para testar o número de grupos. Em seguida, executa-se o algoritmo k-means (FORGY, 1965), variando o número de grupos no intervalo sugerido e usando como centros iniciais os centros identificados no agrupamento hierárquico. Usar tais centros normalmente reduz o impacto de ruídos e requer menos tempo de iteração (LU et al., 2008). Após isso, seleciona-se o número de grupos mais adequado por meio das medidas de variação intragrupos (ANDERBERG, 1973) e índice de Silhouette (ROUSSEEUW, 1987).

A variação intragrupos mede as diferenças entre os trabalhadores e o centro do grupo a que eles pertencem. Quanto menor a variação intragrupos, melhor o agrupamento. Isso indica que os trabalhadores agrupados no mesmo grupo apresentam valores semelhantes para as métricas de engajamento e que o centro do grupo representa o grupo adequadamente. O índice de Silhouette, por sua vez, mede o quão bem separados e coesos os grupos são. Este

índice varia de -1, indicando um agrupamento muito ruim, e 1, indicando um agrupamento excelente. Struyf, Hubert e Rousseeuw (1997) propuseram a seguinte regra (*rule of thumb*) para interpretação do índice de Silhouette: índice entre 0,71 e 1, indica que uma forte estrutura de agrupamento foi encontrada; entre 0,51 e 0,70, indica que uma estrutura razoável foi encontrada; entre 0,26 e 0,50, indica que a estrutura de agrupamento é fraca e pode ser artificial, e, portanto, recomenda-se que os métodos adicionais de análise sejam testados; inferior ou igual a 0,25, indica que nenhuma estrutura substancial foi encontrada.

3.3 Materiais e Métodos de Avaliação

A análise das métricas propostas é realizada usando bases de dados obtidas de projetos reais de computação por humanos. As bases de dados de tais projetos são descritas na próxima seção. Em seguida, os métodos utilizados na avaliação são detalhados.

3.3.1 Descrição dos Projetos Estudados

O estudo do engajamento de trabalhadores em projetos de computação por humanos impõe diversos requisitos em termos de bases de dados. Em cada projeto, é necessária a observação das ações dos trabalhadores ao longo do tempo. Ou seja, é preciso existir um histórico de execução de tarefas. Esse histórico precisa ser longo o suficiente para se analisar a existência de padrões de engajamento de curto prazo, como a atividade diária, e o engajamento de longo prazo, como a variação na periodicidade. Isso significa a disponibilidade de dados de execuções de tarefas no projeto ao longo de semanas ou meses. Por se tratar de um histórico, é necessário que cada atuação do trabalhador no projeto esteja associada à informação do tempo de ocorrência (*timestamp*).

Foram obtidas cinco bases de dados com essas características: Galaxy Zoo (LINTOTT et al., 2008), The Milky Way Project (SIMPSON et al., 2012), Análise de Sentimentos[1], Sun4All e Cell Spotting (LOSTAL et al., 2013). Um sumário estatístico dos dados disponíveis é apresentado na Tabela 3.1. Todas as bases de dados consistem em eventos de execução de tarefas. Tem-se a informação do trabalhador que gerou cada evento e do instante de tempo em que

[1]Os dados desse projeto foram disponibilizados pela empresa CrowdFlower. Para mais informações sobre essa disponibilização, consulte a publicação de lançamento na URL http://www.crowdscale.org/shared-task/sentiment-analysis-judgment-data, acessada pela última vez em 09 de outubro de 2015.

o evento ocorreu. Os parágrafos seguintes detalham essas informações apresentando cada projeto.

Tabela 3.1: Resumo estatístico das bases de dados dos projetos analisados no estudo de engajamento.

Projeto	Duração (dias)	#Trabalhadores	#Eventos
Galaxy Zoo	840	86.413	9.667.586
The Milky Way Project	670	23.889	643.408
Análise de Sentimentos	18	1.960	569.375
Sun4All	305	116	4.328
Cell Spotting	492	1.103	94.137

O projeto Galaxy Zoo consiste em tarefas de classificação de galáxias (LINTOTT et al., 2008). É um projeto de ciência cidadã lançado em julho de 2007. Desde então, ele foi redesenhado e relançado outras vezes. Neste documento são utilizados dados da terceira iteração do Galaxy Zoo, chamada Galaxy Zoo Hubble[2]. Ela foi lançada em abril de 2010 e funcionou até setembro de 2012. A base de dados consiste em 9.667.586 eventos de execuções de tarefas gerados por 86.413 trabalhadores diferentes ao longo de 840 dias.

O projeto The Milky Way Project consiste em tarefas de análise do formato de galáxias (SIMPSON et al., 2012). É um projeto de ciência cidadã lançado em dezembro de 2010 e que permaneceu em operação até setembro de 2012. A base de dados consiste em 643.468 eventos de execuções de tarefas gerados por 23.889 trabalhadores diferentes ao longo de 670 dias.

A base de dados Análise de Sentimentos consiste em tarefas de julgamentos da condição climática relatada em *tweets*[3]. Cada tarefa apresenta ao trabalhador um *tweet* e ele responde se a informação sobre o clima constante no *tweet* é "negativa", "neutra (o autor do *tweet* está apenas compartilhando informação)", "positiva", "O *tweet* não é relacionado à condição do clima", ou "não sei responder". A base de dados consiste em 569.375 eventos de execuções de tarefas gerados por 1.960 trabalhadores diferentes ao longo de 18 dias. Em razão da curta duração, esse projeto só é utilizado em análises de engajamento de curto prazo.

O projeto Sun4All consiste em tarefas de contagem de manchas solares. Cada tarefa apresenta ao trabalhador uma imagem do sol e, usando uma interface gráfica, o trabalhador

[2]Mais informações sobre o projeto estão disponíveis na página oficial http://zoo3.galaxyzoo.org, último acesso 1 de outubro de 2015.

[3]*Tweets* são mensagens de texto de até 140 caracteres compartilhadas na rede social twitter.com.

marca as manchas sobre a imagem e o projeto informa o número de manchas identificadas. A base de dados consiste em 4.328 eventos de execuções de tarefas gerados por 116 trabalhadores diferentes ao longo de 305 dias.

O projeto Cell Spotting consiste em tarefas contagem de células mortas (LOSTAL et al., 2013). Cada tarefa apresenta ao trabalhador uma imagem de uma lâmina contendo diversas células e, usando uma interface gráfica, o trabalhador marca sobre a imagem as células mortas e o projeto informa o número de células mortas identificadas. A base de dados consiste em 94.137 eventos de execuções de tarefas gerados por 1.103 trabalhadores diferentes ao longo de 492 dias.

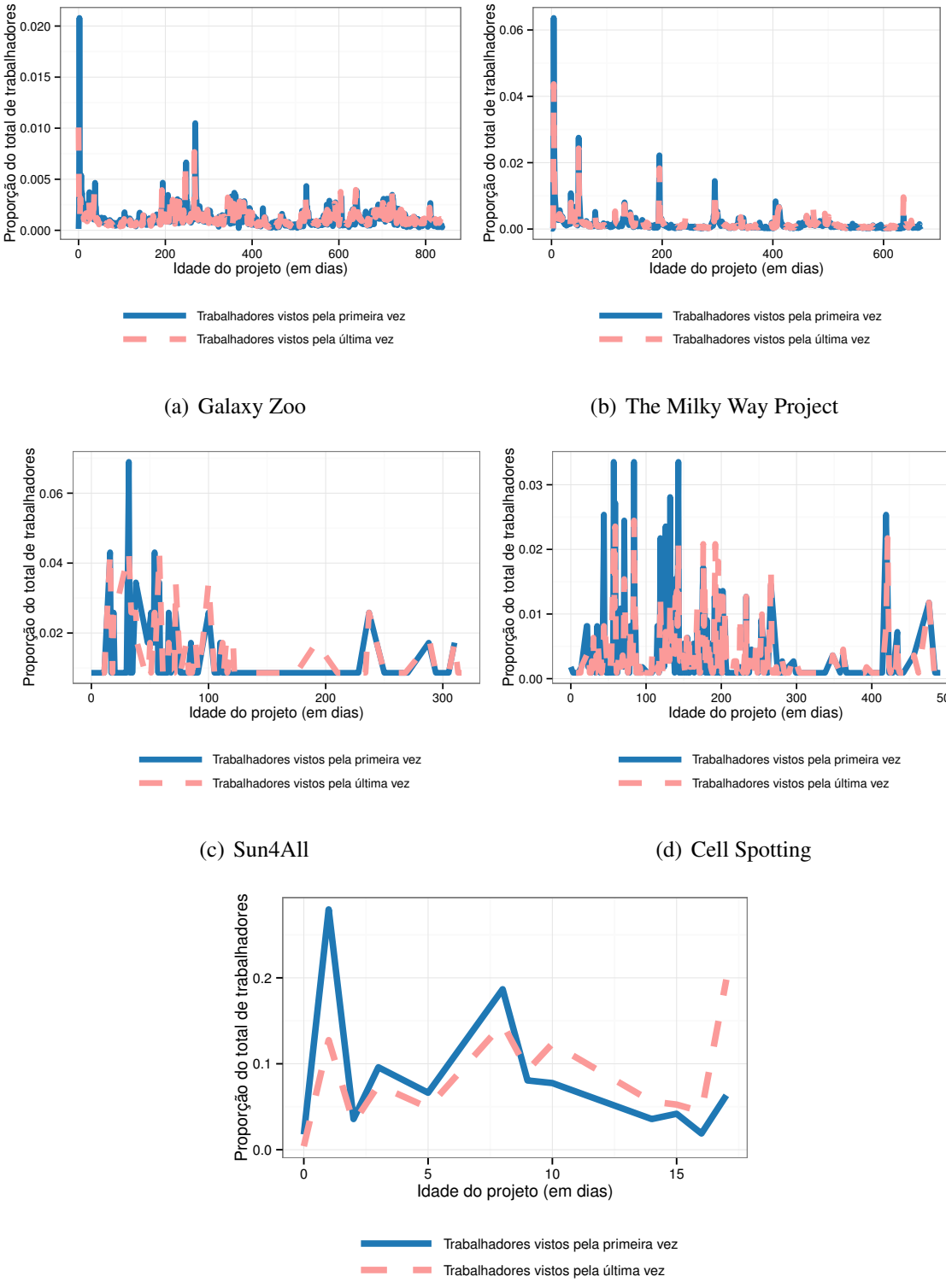
Tratamento dos Trabalhadores que Chegaram ao Projeto Tardiamente

As bases de dados descrevem a dinâmica natural dos projetos. Por essa dinâmica, trabalhadores podem se juntar ao projeto em qualquer momento dentro do período de tempo em que o projeto esteja em operação. Do mesmo modo, eles também podem apresentar algumas atividades e não retornar ao projeto para executar mais tarefas. Essa dinâmica de trabalhadores sendo vistos pela primeira vez e pela última vez nos projetos é apresentada na Figura 3.2. Naturalmente, há grande incerteza sobre o comportamento de trabalhadores que se juntaram ao projeto quando o mesmo estava próximo de ser concluído, pois tais trabalhadores não foram observados por tempo suficiente para se detectar um eventual engajamento de longo prazo.

Para mitigar o efeito desses trabalhadores que chegaram ao projeto tardiamente, foram considerados nas análises apenas trabalhadores que chegaram durante os três primeiros quartos do tempo total de duração do projeto. Esse tratamento resultou na remoção de 2.413 (10%) trabalhadores no projeto The Milky Way Project, 82 (17%) trabalhadores no projeto Cell Spotting, 13 (12%) trabalhadores no projeto Sun4All, 20.182 (23%) trabalhadores no projeto Galaxy Zoo e 315 (16%) trabalhadores no projeto Análise de Sentimentos.

Identificação de Sessões de Trabalho

No estudo do engajamento, é necessário ter acesso à informação das sessões de trabalho de cada trabalhador. Entretanto, as bases de dados utilizadas não delimitam tais sessões. A abordagem utilizada para resolver esse problema foi inferir as sessões a partir do histórico



(a) Galaxy Zoo

(b) The Milky Way Project

(c) Sun4All

(d) Cell Spotting

(e) Análise de Sentimentos

Figura 3.2: Distribuição da proporção de trabalhadores vistos pela primeira vez e vistos pela última vez ao longo dos dias em que o projeto permaneceu em execução. Mostram-se esses comportamentos nos projetos Galaxy Zoo, The Milky Way Project, Sun4All, Cell Spotting e Análise de Sentimentos.

de execução de tarefas. Para cada trabalhador, o histórico marca cada evento de execução de tarefa, como exemplificado na Figura 3.3. Usando a informação do instante de tempo em que esses eventos ocorreram, são calculados os intervalos entre cada execução de duas tarefas sequenciais. Dados esses intervalos, utilizou-se a metodologia de extração de sessões baseada em limiar (ARLITT, 2000) para formar as sessões de trabalho de cada trabalhador. A ideia principal dessa metodologia é, dado um limiar, definir se duas execuções de tarefas consecutivas estão em uma mesma sessão ou se estão em sessões diferentes.

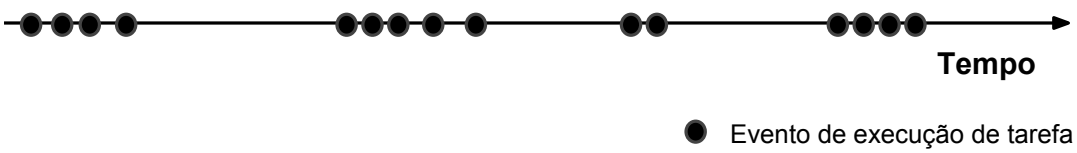


Figura 3.3: Linha do tempo com eventos de execuções de tarefas por um trabalhador em um projeto.

Para encontrar o limiar adequado para cada trabalhador, utiliza-se o método proposto por Mehrzadi e Feitelson (2012). Esse método considera que a distribuição dos intervalos entre cada par de execuções seguidas forma um histograma bimodal no qual um grupo é formado pelos curtos intervalos entre cada par de execuções e outro grupo é formado pelos longos intervalos entre cada par de execuções. O método identifica um limiar que divide esses dois grupos. Neste trabalho, um curto intervalo entre execuções de tarefas representa o tempo durante o qual o trabalhador está pensando a solução para tarefa, o que indica que ele não terminou a sessão de trabalho. Um longo intervalo, por sua vez, representa o tempo que é maior do que o trabalhador gastaria para solucionar uma tarefa, indicando que o trabalhador terminou a sessão de trabalho. Para exemplificar, a Figura 3.4 mostra o histograma dos intervalos de tempo entre execuções de tarefas de um trabalhador no Galaxy Zoo. Essa figura também indica o limiar de 1.024 segundos (17,07 minutos) identificado para esse trabalhador.

É importante ressaltar que existem diversos métodos de identificação de sessões de trabalho. Existem inclusive abordagens que usam um limiar fixo para todos os trabalhadores (MAO; KAMAR; HORVITZ, 2013; GEIGER; HALFAKER, 2013). No entanto, até onde se sabe, o método escolhido para ser utilizado neste trabalho é o único proposto para se identificar um limiar para cada ser humano que atua no sistema (no caso desta pesquisa, para cada

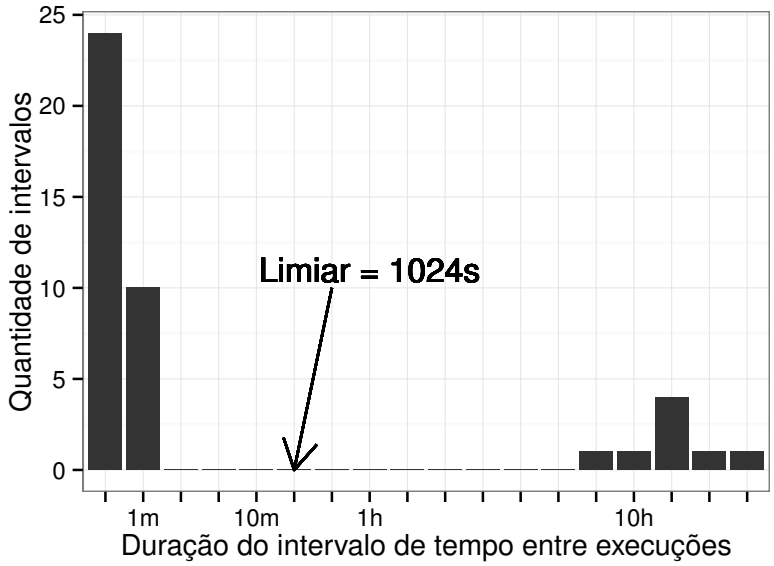


Figura 3.4: Exemplo do limiar identificado para um trabalhador no projeto Galaxy Zoo. Histograma com células em dimensão logarítmica de base 2.

trabalhador nos projetos). Isso está alinhado com um dos objetivos do trabalho proposto neste documento que é analisar as diferenças entre trabalhadores em cada projeto. Como mostrado na Tabela 3.2, a maioria dos trabalhadores nas bases de dados apresenta o limiar de 1.024 segundos. Os demais se distribuem em limiares de 512 segundos, 2.048 segundos e 4.096 segundos.

Tabela 3.2: Distribuição dos trabalhadores em diferentes valores de limiares de delimitação das sessões de trabalho.

Limiar	Galaxy Zoo	The Milky Way Project	Cell Spotting	Análise de Sentimentos	Sun4All
512s	7%	5%	9%	6%	6%
1.024s	**85%**	**82%**	**77%**	**70%**	**81%**
2.048s	5%	9%	12%	14%	7%
4.096s	2%	4%	2%	11%	6%

3.3.2 Método de Caracterização de Semelhanças e Diferenças entre Trabalhadores

É comum, em sistemas Web, se analisar a concentração de trabalhadores que são ativos em apenas um dia em comparação com aqueles que permanecem ativos por mais dias. No

contexto de computação por humanos, os primeiros são chamados *transientes* enquanto os segundos são chamados *regulares*. Trabalhadores transientes podem ser apenas curiosos, ou pessoas que concentram toda sua contribuição em um único dia. Os trabalhadores regulares, por sua vez, podem ser pessoas dispostas a se manterem engajadas no projeto por mais tempo e se manterem participativas até o projeto ser concluído. A caracterização realizada analisa a concentração e a importância dessas classes de trabalhadores.

A análise das semelhanças e diferenças entre os trabalhadores em um dado projeto é realizada considerando as funções de distribuição acumulada (FDAs) das métricas de engajamento. Essas distribuições informam o quão os trabalhadores se concentram em uma determinada faixa de valores para uma dada métrica em análise. Em algumas situações, mostra-se importante investigar em que medida duas FDAs são semelhantes. Por exemplo, verificar se a distribuição dos trabalhadores pela métrica taxa de atividade é igual à distribuição dos trabalhadores pela métrica duração relativa da atividade. Outro caso de interesse é aquele em que se deseja saber se a distribuição dos trabalhadores em uma métrica em um dado projeto é semelhante à distribuição dos trabalhadores pela mesma métrica em outro projeto. Essas comparações de FDAs são realizadas por meio da estatística D do teste Two-sample Kolmogorov-Smirnov (SMIRNOV, 1939). Essa estatística indica a distância entre duas FDAs. Na análise dos resultados, para dizer que duas distribuições são iguais, considera-se o grau de significância de $0,05$.

As métricas de engajamento também podem ser utilizadas na identificação de perfis de engajamento que descrevem padrões naturais com que os trabalhadores se engajam em cada projeto. Os perfis são identificados usando o algoritmo de agrupamento k-means e a qualidade do agrupamento é avaliada usando as métricas Índice de Silhouette e Variação Intra Grupo. Esse método é empregado na análise das semelhanças e diferenças do engajamento dos trabalhadores. A fim de compreender os diferentes grupos identificados, analisa-se: (i) os centróides que representam os grupos; e (ii) a relação entre cada par de métricas de engajamento dos trabalhadores em cada grupo. Analisa-se também a forma como os grupos diferem em termos do número de trabalhadores e do total de tempo de computação agregado ao projeto.

3.3.3 Método de Caracterização e Análise de Relações entre Métricas

Via de regra, nas análises de relações entre variáveis, utiliza-se o coeficiente de correlação de Spearman (SPEARMAN, 1904), definido como ρ. Trata-se de uma medida de correlação não paramétrica e que detecta tanto relações lineares quanto não-lineares. O coeficiente ρ assume valores no intervalo entre -1 e 1. Um valor positivo indica que o relacionamento entre as duas variáveis em análise possui a mesma direção, por exemplo, se o valor de uma variável cresce o valor da outra também cresce. Por outro lado, um valor negativo indica que o relacionamento entre as duas variáveis em análise é inverso, ou seja, se o valor de uma variável cresce o valor da outra decresce. Quanto mais próximo de -1 ou 1, mais forte é a relação entre as variáveis e, quanto mais próximo de 0, mais insignificante é essa relação. Na análise dos resultados de correlação, utiliza-se um nível de significância estatística de 0,05. Assim, a correlação é dita significante quando o intervalo de confiança de 95% não inclui o valor 0 ou um teste estatístico diz que o valor é significativamente diferente de 0.

3.4 Apresentação e Análise dos Resultados

Esta seção reúne os resultados obtidos. Um primeiro resultado a ser discutido na caracterização do engajamento é a participação de trabalhadores transientes e regulares. A Tabela 3.3 mostra a concentração de trabalhadores regulares e transientes nos projetos analisados e a contribuição deles em termos do total de tempo dedicado ao projeto.

3.4.1 Transientes e Regulares

No que se refere à concentração dos trabalhadores, os resultados destacados na Tabela 3.3 indicam que a *maioria dos trabalhadores que chegam aos projetos são transientes e que a minoria é regular*. Trabalhadores regulares consistem em apenas 28% no projeto The Milky Way Project, 35% no projeto Sun4All, 36% no projeto Galaxy Zoo e 42% no projeto Cell Spotting. Essa baixa concentração de trabalhadores regulares é uma evidência de que *apenas uma pequena parte dos trabalhadores que atuam nesse tipo de projeto apresentam um engajamento de longo prazo*.

No que se refere à importância, quando se analisa a contribuição dos trabalhadores para

Tabela 3.3: Concentração e importância dos trabalhadores Transientes e Regulares.

Projeto	Nº de trabalhadores		Tempo dedicado (horas)	
	Transientes	Regulares	Transientes	Regulares
Galaxy Zoo	42.684 (64%)	23.547 (36%)	8.680 (14%)	51.467 (86%)
The Milky Way Project	15.323 (72%)	6.093 (28%)	2.889 (16%)	14.650 (84%)
Cell Spotting	595 (58%)	425 (42%)	122 (10%)	1.126 (90%)
Sun4All	66 (65%)	36 (35%)	4 (6%)	70 (94%)
Análise de sentimentos	978 (59%)	667(41%)	73 (7%)	2.194(93%)

o total de tempo dedicado à execução de todas as tarefas no projeto, observa-se que a quase totalidade desse tempo também foi dedicado por trabalhadores regulares. Os trabalhadores regulares são responsáveis por 94% do tempo de computação dedicado ao projeto Sun4All, 90% do tempo dedicado ao projeto Cell Spotting, 86% do tempo dedicado ao projeto Galaxy Zoo e 84% do total de tempo dedicado ao projeto The Milky Way Project. Esse resultado indica que *os trabalhadores regulares são fundamentais para os projetos* uma vez que eles apresentam maior contribuição em termos do total de tempo de computação agregado ao projeto.

Considerando a concentração e a importância, observa-se que os trabalhadores regulares, embora sejam a minoria, eles são os que mais contribuem para os projetos. Ou seja, não é suficiente para o sucesso de um projeto existir uma multidão de trabalhadores que exibem apenas um engajamento de curto prazo. Ao final do projeto, os trabalhadores que se revelam mais importantes são aqueles que compõem a minoria que persistiu no projeto de forma mais duradoura. Dessa forma, entender o comportamento de tais trabalhadores se mostra crucial para se compreender o desempenho de projetos de computação por humanos.

3.4.2 Distribuições do Engajamento

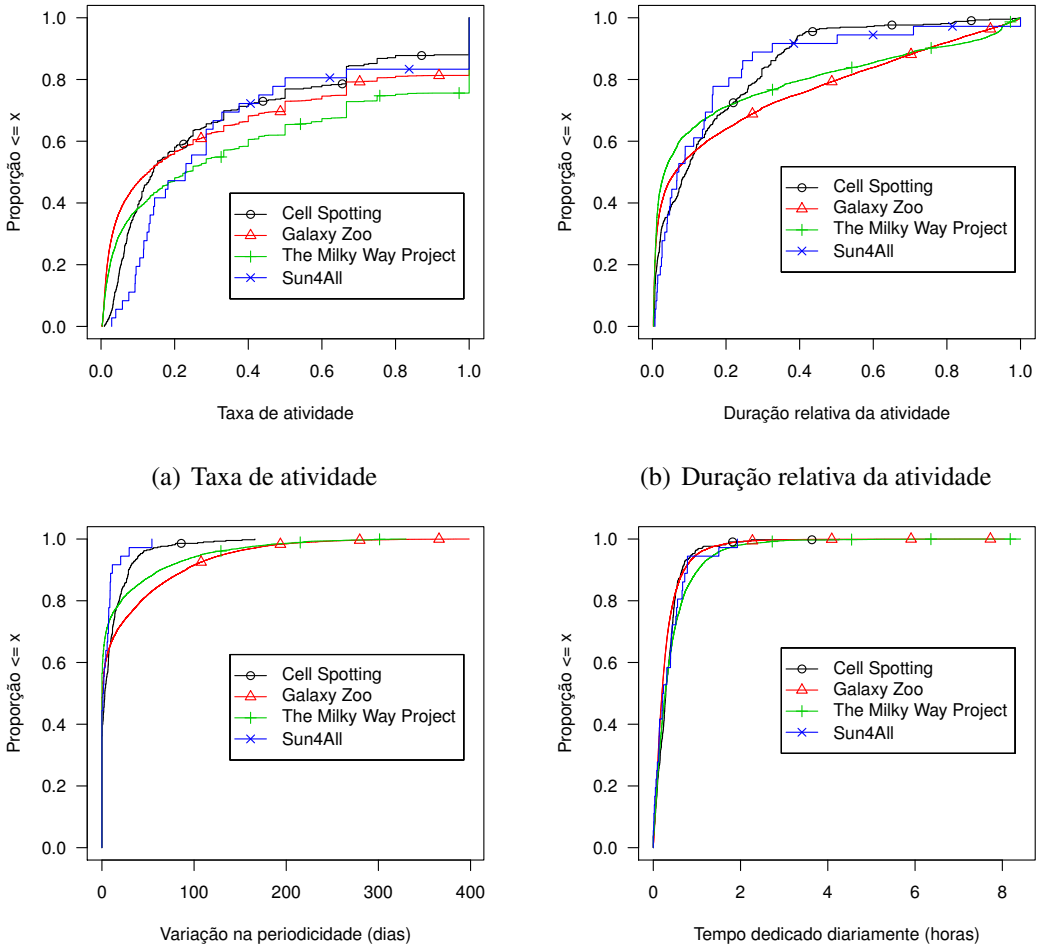
Um dos primeiros passos para se entender melhor o engajamento dos trabalhadores regulares é ter uma perspectiva clara sobre como são as distribuições do engajamento desses trabalhadores em cada projeto. Tais distribuições são apresentadas na Figura 3.5 que destaca as distribuições por projeto e por métrica de engajamento: taxa de atividade (Fig 3.5(a)), duração relativa da atividade (Fig 3.5(b)), variação na periodicidade (Fig 3.5(c)) e tempo dedicado diariamente (Fig 3.5(d)).

No resultado para a métrica taxa de atividade (Fig 3.5(a)), observa-se que os trabalhado-

res regulares não são semelhantes entre si. Há certa concentração em valores baixos e uma dispersão em valores maiores. Por exemplo, considerando a mediana (0,5 quantil) observa-se certa concentração dos trabalhadores em taxas de atividades em valores baixos, no intervalo de 0 e 0,23. Galaxy Zoo é o projeto que apresenta o menor valor mediano, exibindo uma taxa de atividade mediana de 0,13. Isso indica que até 50% dos trabalhadores permanecem ativos em 13% do total de dias decorridos entre seu primeiro e último dia ativo no projeto. Isso é equivalente a acessarem o projeto quase uma vez por semana. The Milky Way Project, por sua vez, é o projeto em que os trabalhadores tendem a ser mais ativos. Nesse projeto até 50% dos trabalhadores apresentaram uma taxa de atividade de até 0,23. Os demais projetos apresentam medianas intermediarias. No geral, esses resultados indicam uma *baixa atividade dos trabalhadores durante o período de tempo que eles ficam no projeto*.

Quando se analisa a distribuição dos trabalhadores segundo a métrica duração relativa da atividade (Fig 3.5(b)), observa-se grande concentração de trabalhadores com valores muito baixos. Em todos os projetos, 50% dos trabalhadores apresentam duração relativa da atividade de até 0,09. Esse resultado caracteriza que *os trabalhadores permanecem ligados ao projeto durante um período de tempo muito pequeno em relação ao tamanho do período de tempo que eles poderiam permanecer*, indicando que eles normalmente deixam de retornar ao projeto para executar tarefas muito antes do projeto ser concluído. Isso é ainda mais intenso nos projetos Sun4All e Cell Spotting, nos quais 90% dos trabalhadores apresentam uma duração relativa de atividade menor ou igual a 0,30 e 0,36 respectivamente.

Quando à variação na periodicidade (Fig 3.5(c)), os valores máximos observados foram de 399 dias no projeto Galaxy Zoo, 330 dias no projeto The Milky Way Project, 166 dias no projeto Cell Spotting e 54 dias no projeto Sun4All. Entretanto, em todos os projetos se observa uma concentração da variação em valores baixos. A mediana da variação na periodicidade é de 0 dias no projeto The Milky Way Project, de 0,5 dias no projeto Galaxy Zoo, 1,48 dias no projeto Sun4All e 3,12 dias no projeto Cell Spotting. Esses resultados indicam que, embora existam trabalhadores que apresentam grande variação na periodicidade, no geral, *a maioria dos trabalhadores retornam ao projeto com uma periodicidade que varia em poucos dias*. As distribuições dos trabalhadores pela métrica variação na periodicidade nos projetos são bastante próximas até o 75 percentil. Após esse limiar, tornam-se distantes dependendo do projeto.



(a) Taxa de atividade

(b) Duração relativa da atividade

(c) Variação na periodicidade

(d) Tempo dedicado diariamente

Figura 3.5: Funções de distribuição acumulada (FDAs) dos trabalhadores nos projetos Galaxy Zoo, The Milky Way Project, Sun4All e Cell Spotting de acordo com as métrica de engajamento: (a) Taxa de atividade, (b) Duração relativa da atividade, (c) Variação na periodicidade e (d) Tempo dedicado diariamente.

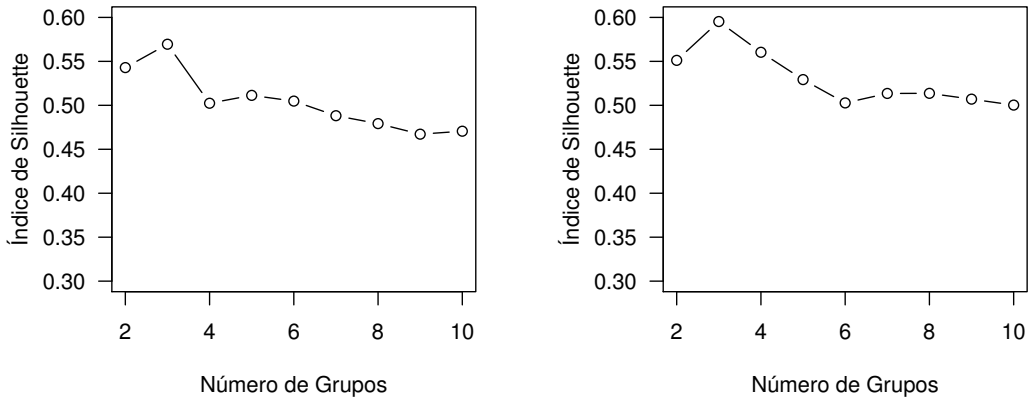
Finalmente, quando se considera a métrica tempo dedicado diariamente (Fig 3.5(d)), observa-se grande semelhança entre as distribuições dos trabalhadores nos diversos projetos estudados. A mediana das distribuições fica entre o menor valor de 0,20 horas, no projeto Galaxy Zoo, e o maior valor de 0,29 horas, no projeto Cell Spotting. De uma forma geral, 90% dos trabalhadores dedicam-se aos projetos durante um tempo diário de até 1 hora.

Notadamente, as distribuições em cada métrica são parecidas apesar das diferenças dos projetos em termos do tipo de tarefas e do público de trabalhadores que eles atraem. As distâncias entre as distribuições são baixas. Em todas as comparações possíveis, os resultados de distância são sempre menores ou iguais a 0,35. Em alguns casos nem há evidência estatística de que as distribuições sejam oriundas de populações diferentes. Isso ocorre na distância entre os projetos Cell spotting e Sun4All na métrica duração relativa da atividade.

3.4.3 Perfis de Engajamento

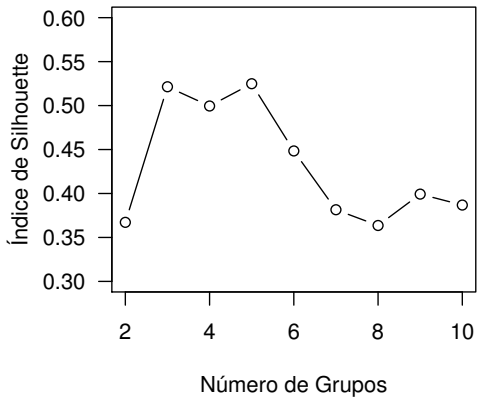
Além de analisar o comportamento dos trabalhadores em cada métrica de engajamento independentemente, é importante entender o comportamento dos trabalhadores quando todas as métricas são consideradas ao mesmo tempo. Isso é feito por meio do agrupamento dos trabalhadores pelos valores de engajamento que eles apresentam. Nesse sentido, agrupou-se os trabalhadores que apresentam características de engajamento semelhantes. A análise da qualidade de agrupamentos considerando diferentes números de grupos é apresentada na Figura 3.6, referente ao Índice de Silhouette, e na Figura 3.7, referente à variação intragrupos.

Identificou-se que o agrupamento em cinco 5 grupos se mostra adequado em todos os projetos. Nos projetos Galaxy Zoo, The Milky Way Project e Sun4All o índice de Silhouette no agrupamento com esse número de grupos é superior a 0,51. Esse valor indica que uma estrutura de agrupamento razoável foi encontrada nos dados. Além disso, nesses projetos, também se observa uma queda acentuada na variação intragrupo até o número de 5 grupos. Ou seja, não há ganho aparente em se utilizar um número de grupos maior ou menor que 5. Já no projeto Cell Spotting, o índice de Silhouette para 5 grupos é aproximadamente 0,45 o que indica que pode existir uma estrutura de agrupamento e que um método de avaliação alternativo deve ser considerado. Assim, observou-se que, também nesse projeto, ocorre uma queda significativa na variação intragrupo até o número de 5 grupos. Dessa forma, também neste projeto, a estrutura de 5 grupos se mostra adequada. Cabe agora compreender

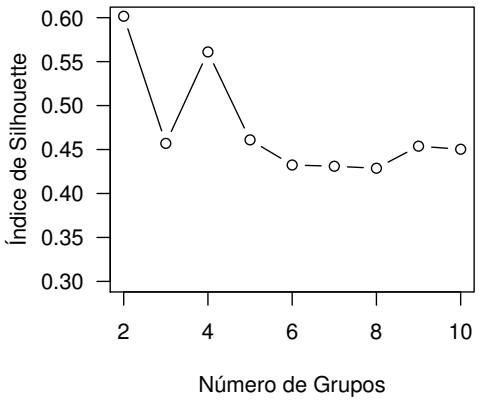


(a) Galaxy Zoo

(b) The Milky Way Project



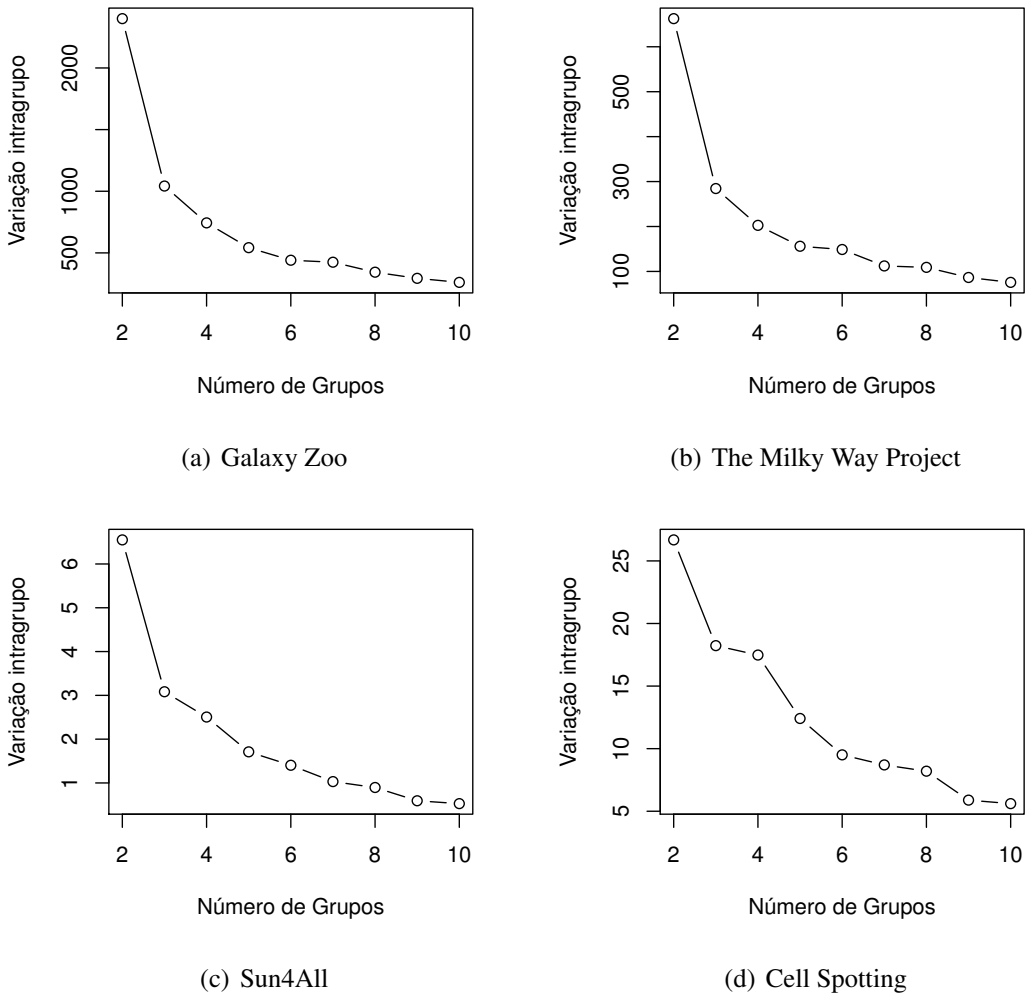
(c) Sun4All



(d) Cell Spotting

Figura 3.6: Índice de Silhouette em agrupamentos gerados pelo algoritmo k-means quando o número de grupos é variado. Mostram-se resultados obtidos nos projetos (a) Galaxy Zoo, (b) The Milky Way Project, (c) Sun4All e (d) Cell Spotting.

os significados dos grupos identificados.



(a) Galaxy Zoo

(b) The Milky Way Project

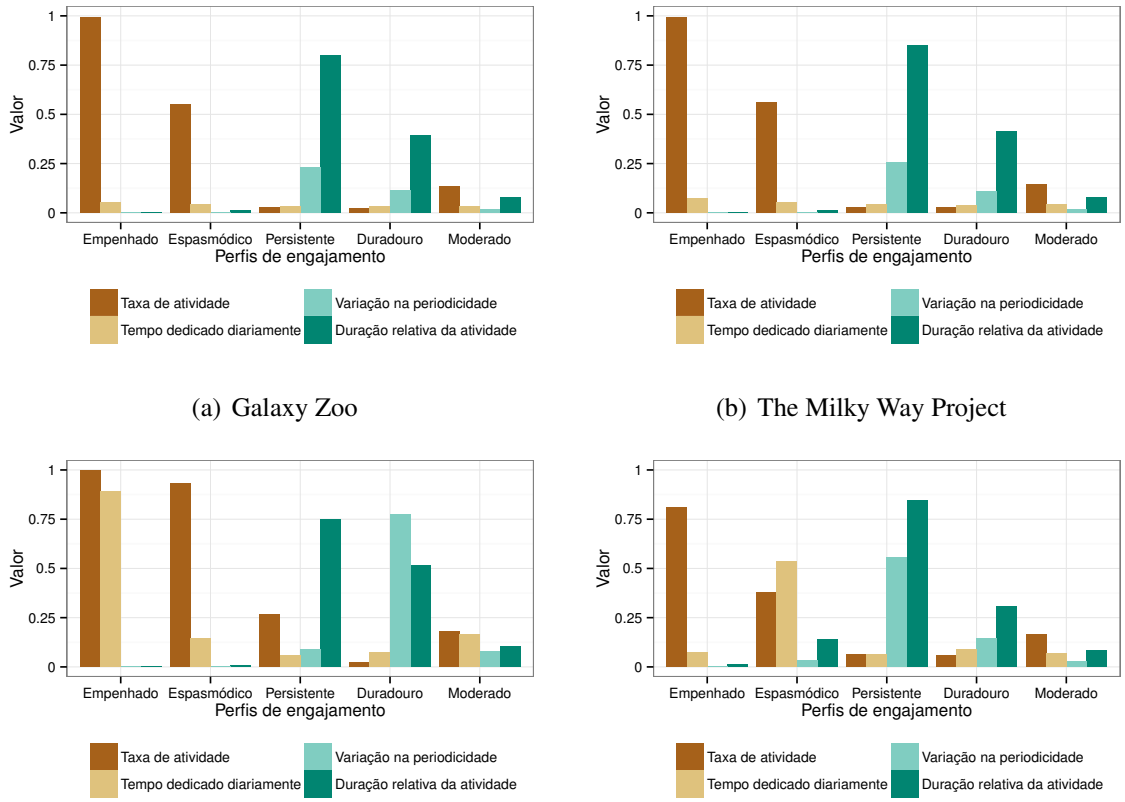
(c) Sun4All

(d) Cell Spotting

Figura 3.7: Variação intragrupos em agrupamentos gerados pelo algoritmo k-means quando o número de grupos é variado. Mostram-se resultados obtidos nos projetos (a) Galaxy Zoo, (b) The Milky Way Project, (c) Sun4All e (d) Cell Spotting.

Ao analisar o comportamento dos trabalhadores em cada grupo, foram estabelecidos rótulos para os grupos a fim de colocar em perspectiva suas principais características de engajamento. Os cinco grupos foram rotulados como: empenhado, espasmódico, persistente, duradouro, e moderado. Dessa forma, cada grupo representa um perfil de engajamento. Na definição dos rótulos, considerou-se: (i) os centróides que representam os grupos em termos das métricas de engajamento (Fig 3.8); e (ii) inter-relações existentes entre cada par de métricas de e engajamento em cada grupo (Tabela 3.4). A Figura 3.8 mostra os valo-

res das métricas taxa de atividade, variação na periodicidade, tempo dedicado diariamente e duração relativa da atividade exibidos pelo centróide de cada perfil de engajamento. A Tabela 3.4, por sua vez, mostra as correlações entre essas métricas no conjunto de trabalhadores que compõem cada perfil. Nessa tabela, há valores não definidos no projeto Sun4All. Isso ocorre porque apenas o perfil de engajamento moderado agregou um número de trabalhadores grande o bastante (> 5) para que as correlações fossem calculadas. Nos próximos parágrafos os perfis são apresentados à luz desses resultados.



(a) Galaxy Zoo

(b) The Milky Way Project

(c) Sun4All

(d) Cell Spotting

Figura 3.8: Centróides dos perfis de engajamento dos trabalhadores em termos das métricas taxa de atividade, duração relativa da atividade, tempo dedicado diariamente e variação na periodicidade. Mostram-se resultados obtidos nos projetos: (a) Galaxy Zoo, (b) The Milky Way Project, (c) Sun4All e (d) Cell Spotting.

Trabalhadores Empenhados. Os trabalhadores que apresentam um perfil de engajamento empenhado têm maior taxa de atividade e menor duração relativa da atividade em comparação com os trabalhadores que exibem os outros perfis (Fig 3.8). Essas métricas indicam que os trabalhadores neste perfil trabalham quase diariamente quando eles chegam ao

Tabela 3.4: Correlação de Spearman entre cada par de métricas de engajamento nos perfis de trabalhadores regulares.

Galaxy Zoo

Par	Empenhado $N = 4.572$	Espasmódico $N = 3.611$	Persistente $N = 3.783$	Duradouro $N = 4.250$	Moderado $N = 7.331$
$\rho(a, r)$	-0,30*	**-0,45***	0,15*	-0,23*	**-0,76***
$\rho(a, v)$	**-0,99***	-0,31*	-0,26	0,27*	-0,12*
$\rho(a, d)$	-0,10*	0,03	0,33*	0,30*	0,19*
$\rho(r, v)$	0,30*	**0,66***	-0,12*	0,00	**0,43***
$\rho(r, d)$	0,07*	0,17*	0,08*	0,02	-0,05*
$\rho(v, d)$	0,10*	0,26*	-0,01	0,16*	0,16*

The Milky Way Project

Par	Empenhado $N = 1.535$	Espasmódico $N = 1.060$	Persistente $N = 817$	Duradouro $N = 844$	Moderado $N = 1.837$
$\rho(a, r)$	-0,24*	-0,38*	-0,14*	-0,26*	**-0,74***
$\rho(a, v)$	**-0,99***	-0,22*	0,06	0,39*	-0,13*
$\rho(a, d)$	-0,07*	-0,05	**0,43***	0,37*	0,14*
$\rho(r, v)$	0,24*	**0,59***	-0,13*	-0,04	**0,44***
$\rho(r, d)$	0,14*	0,23*	-0,09*	0,02	0,01
$\rho(v, d)$	0,07*	0,29*	0,19*	0,31*	0,21*

Cell Spotting

Par	Empenhado $N = 109$	Espasmódico $N = 7$	Persistente $N = 11$	Duradouro $N = 117$	Moderado $N = 181$
$\rho(a, r)$	**-0,71***	-0,75	0,07	0,01	**-0,62***
$\rho(a, v)$	**-0,47***	**-0,96***	**-0,97***	**-0,41***	-0,12
$\rho(a, d)$	0,09	-0,32	0,52	0,26	0,05
$\rho(r, v)$	**0,69***	**0,82***	0,14	0,23	**0,47***
$\rho(r, d)$	0,01	0,07	-0,06	-0,04	0,13
$\rho(v, d)$	0,04	0,35	-0,39	0,01	0,18*

Sun4All

Par	Empenhado $N = 2$	Espasmódico $N = 5$	Persistente $N = 2$	Duradouro $N = 2$	Moderado $N = 25$
$\rho(a, r)$	–	–	–	–	**-0,66***
$\rho(a, v)$	–	–	–	–	-0,33
$\rho(a, d)$	–	–	–	–	0,07
$\rho(r, v)$	–	–	–	–	**0,70***
$\rho(r, d)$	–	–	–	–	**0,45***
$\rho(v, d)$	–	–	–	–	**0,61***

Nota 1: Correlações entre as métricas de engajamento taxa de atividade (a), duração relativa da atividade (r), tempo dedicado diariamente (d) e variação na periodicidade (v) exibidas pelos N trabalhadores em cada perfil de engajamento.
Nota 2: *Coeficiente de correlação de Spearman ρ significante (p-valor < 0,05).
Nota 3: – Quantidade de dados muito pequena para se calcular a correlação e o p-valor.

projeto, mas eles deixam o projeto em pouco tempo após chegarem. Este perfil de engajamento também apresenta baixa variação na periodicidade. Isto significa que os trabalhadores que apresentam esse perfil de engajamento retornam ao projeto para executar mais tarefas em intervalos de tempo quase iguais. Isso faz com que o tempo de retorno destes trabalhadores seja razoavelmente previsível. Outra característica desse grupo de trabalhadores, que pode ser observada na Tabela 3.4, é que eles apresentam uma forte correlação negativa entre a taxa de atividade e a variação na periodicidade (Galaxy Zoo, $\rho(a, v)$ = -0,99; The Milky Way Project, $\rho(a, v)$ = -0,99; Cell Spotting, $\rho(a, v)$ = -0,47). Essas correlações indicam que quanto mais frequentes são os retornos dos trabalhadores durante o período que eles permanecem nos projetos, menos variáveis são os intervalos de tempo decorridos entre esses retornos.

Trabalhadores Espasmódicos. Este perfil de engajamento se distingue por apresentar uma taxa de atividade relativamente alta e duração da atividade relativamente baixa (Fig 3.8). Este grupo de trabalhadores apresenta uma forte correlação positiva entre duração relativa da atividade e variação na periodicidade (Galaxy Zoo, $\rho(r, v)$ = 0,66; The Milky Way Project, $\rho(r, v)$ = 0,59; Cell Spotting, $\rho(r, v)$ = 0,82). Essas correlações indicam que quanto maior o período de tempo que os trabalhadores permanecem ligados ao projeto, mais irregular é a periodicidade de seu retorno para o projeto dentro deste período. Todas estas características indicam que a contribuição de trabalhadores que exibem este perfil tipicamente ocorre durante um curto período de tempo e com a periodicidade mais irregular quanto maior for a duração desse período.

Trabalhadores Persistentes. Engajamento persistente é caracterizado pela maior duração relativa da atividade, alta variação na periodicidade e uma baixa taxa de atividade (Fig 3.8). Assim, os trabalhadores com um perfil de engajamento persistente permanecem ligados ao projeto por um longo intervalo de tempo, mas são ativos apenas em alguns dias dentro desse intervalo. Considerando essas métricas de engajamento, o engajamento persistente pode ser visto como o oposto do engajamento empenhado. Não se observou nenhuma correlação nesse grupo de trabalhadores que seja consistente nos diversos projetos estudados. Entretanto, algumas relações isoladas podem ser observadas. No projeto The Milky Way Project, trabalhadores persistentes que exibem maior taxa de atividade também tendem a exibir maior tempo dedicado diariamente ($\rho(a, d)$ = 0,43). Já no projeto Cell Spotting, essa correlação entre taxa de atividade e tempo dedicado diariamente é alta ($\rho(a, d)$ = 0,52), mas

não significante. Dessa forma, observa-se que o comportamento dos trabalhadores dentro do perfil é bastante variável com o projeto.

Trabalhadores Duradouros. Este é o perfil engajamento dos trabalhadores que exibem elevada duração relativa da atividade e variação na periodicidade (Fig 3.8). Este perfil de trabalhadores apresenta uma taxa de atividade semelhante à taxa que é exibida pelos trabalhadores que permanecem mais tempo no projeto (os de engajamento persistente), mas eles permanecem no projeto durante um período de tempo menor. Também não se observou nenhuma correlação nesse grupo de trabalhadores que seja consistente nos diversos projetos estudados. Uma característica específica do projeto Cell Spotting é que os trabalhadores duradouros que exibem maior taxa de atividade tendem a exibir menor variação na periodicidade ($\rho(a, v)$ = -0,41).

Trabalhadores Moderados. Como mostrado na Figura 3.8, este perfil de engajamento não tem métricas de engajamento particularmente distinguíveis. Assim, em comparação com os outros perfis, trabalhadores moderados apresentam valores intermediários em todas as métricas de engajamento. Uma característica importante do engajamento moderado é uma forte correlação negativa entre a taxa de atividade e duração relativa da atividade (Galaxy Zoo, $\rho(a, r)$ = -0,76; The Milky Way Project, $\rho(a, r)$ = -0,74; Cell Spotting, $\rho(a, r)$ = -0,62; Sun4All, $\rho(a, r)$ = -0,66). Essa correlação indica que o grau de engajamento neste perfil reduz com o aumento da duração do engajamento. Assim, quanto mais dias os trabalhadores retornam ao projeto para executar tarefas, menor é o período total de tempo que eles permanecem ligados ao projeto. Do mesmo modo que os trabalhadores que apresentam um engajamento espasmódico, os trabalhadores que apresentam um engajamento moderado exibem uma correlação positiva entre duração relativa da atividade e variação na periodicidade (Galaxy Zoo, $\rho(r, v)$ = 0,43; The Milky Way Project, $\rho(r, v)$ = 0,44; Cell Spotting, $\rho(r, v)$ = 0,47; Sun4All, $\rho(r, v)$ = 0,70). Isso indica que quanto maior o período de tempo que tais trabalhadores permanecem ligados ao projeto, mais irregular é a periodicidade de seus retornos para o projeto dentro deste período.

Nessa análise de perfis, observa-se a existência de *dois perfis de alto grau de engajamento no curto prazo (empenhados e espasmódicos) e dois perfis de alta duração do engajamento (persistentes e duradouros)*. As características gerais que caracterizam os perfis podem ser observadas em todos os projetos, mas também existem singularidades de alguns projetos

que cabem ser ressaltadas. No projeto Cell Spotting, é notável um maior tempo dedicado diariamente pelos trabalhadores espasmódicos (Fig 3.8(d)). O mesmo ocorre com os trabalhadores que exibem o perfil de engajamento empenhado no projeto Sun4All (Fig 3.8(c)). Nesse último, a alta variação na periodicidade dos trabalhadores duradouros também é uma característica desse perfil apenas nesse projeto. Como já discutido, também existem correlações entre métricas que surgem em um dado perfil e em um dado projeto, mas que não surgem em outros.

Finalmente, mostra-se importante analisar os grupos identificados quanto à concentração e a importância para os projetos. A Tabela 3.5 mostra como os perfis identificados diferem em termos do número de trabalhadores que exibem o perfil e do total de tempo de computação que esses trabalhadores agregaram ao projeto. Nos quatro projetos avaliados, a *maior parte dos trabalhadores apresenta um perfil de engajamento moderado* (31% no projeto Galaxy Zoo, 30% no projeto The Milky Way Project, 43% no projeto Cell Spotting e 69% no projeto Sun4All). Isso indica que esse é o perfil de engajamento típico dos trabalhadores regulares nos projetos. Apenas no projeto Sun4All, tais trabalhadores de engajamento moderado também apresentaram a maior contribuição agregada em termos de tempo de computação dedicado. Nesse projeto, eles contribuíram com 51,46% do tempo de computação. No projeto Galaxy Zoo e The Milky Way Project os trabalhadores que apresentam o perfil de engajamento persistente são os que dedicaram mais tempo ao projeto (46% e 40%, respectivamente). Já no projeto Cell Spotting, foram os trabalhadores de perfil duradouro que dedicaram mais tempo (32% do total).

3.5 Considerações Finais

Neste capítulo, discutiu-se o uso do conceito de engajamento de seres humanos na caracterização da oferta de poder cognitivo pelos trabalhadores em projetos de computação por humanos. Têm-se quatro principais contribuições neste estudo. A primeira contribuição é a contextualização do conceito de engajamento em projetos de computação por humanos. A segunda contribuição é a proposta de quatro métricas para analisar o engajamento de trabalhadores em computação por humanos, são elas: taxa de atividade, tempo dedicado diariamente, duração relativa da atividade e variação na periodicidade. A terceira contribuição

Tabela 3.5: Importância dos perfis em termos do número de trabalhadores e tempo dedicado.

Perfil	Galaxy Zoo		The Milky Way Project	
	Trabalhadores	**Tempo dedicado**	**Trabalhadores**	**Tempo dedicado**
Empenhado	4.572 (19,42%)	4.857,49 (9.44%)	1.535 (25,19%)	2.030,26 (13,86%)
Espasmódico	3.611 (15,34%)	6.061,40 (11,78%)	1.060 (17,40%)	1.912,05 (13,05%)
Persistente	3.783 (16,07%)	**23.757,64 (46,16%)**	817 (13,41%)	**5.846,58 (39,91%)**
Duradouro	4.250 (18,05%)	8.168,95 (15,87%)	844 (13,85%)	2.273,10 (15,52%)
Moderado	**7.331 (31,13%)**	8.621,64 (16,75%)	**1.837 (30,15%)**	2.588,28 (17,67%)
soma	23.547 (100%)	51.467,12 (100%)	6.093 (100%)	14.650,27 (100%)

	CellSpotting		Sun4All	
	Trabalhadores	**Tempo dedicado**	**Trabalhadores**	**Tempo dedicado**
Empenhado	109 (25,65%)	154,32 (13,70%)	2 (5,56%)	9,87 (14,00%)
Espasmódico	7 (1,65%)	250,44 (22,24%)	5 (13,89%)	4,84 (6,87%)
Persistente	11 (2,59%)	91,46 (8,12%)	2 (5,56%)	17,97 (25,52%)
Duradouro	117 (27,53%)	**356,29 (31,63%)**	2 (5,56%)	1,51 (2,14%)
Moderado	**181 (42,59%)**	273,77 (24,31%)	**25 (69,44%)**	**36,24 (51,46%)**
soma	425 (100%)	1126,28 (100%)	36 (100%)	70,43 (100%)

Nota: Para cada projeto, encontram-se destacados em negrito o maior número de trabalhadores em um mesmo perfil e o maior tempo dedicado pelo total de trabalhadores em um mesmo perfil.

é a abordagem de identificação de perfis de engajamento a partir das métricas propostas. Por fim, a quarta contribuição são os comportamentos identificados ao caracterizar a participação dos trabalhadores nos projetos usando a abordagem de engajamento.

As métricas de engajamento e a abordagem de identificação de perfis de engajamento foram utilizadas na análise da oferta de poder cognitivo por trabalhadores em quatro projetos de computação por humanos reais. Os projetos analisados diferem entre si em termos da quantidade de trabalhadores, da quantidade de tarefas, das características das tarefas e do tempo total de duração. Os principais resultados obtidos nos projetos analisados foram:

- A maioria dos trabalhadores que atua nos projetos é transiente, aqueles que atuam em dia e não voltam mais, e a minoria é regular, aqueles que atuam no projeto em mais de um dia;

- Apesar de serem a minoria, os trabalhadores regulares são fundamentais para os projetos, pois apresentam a maior contribuição em termos do total de tempo de computação agregado ao projeto;

- Trabalhadores regulares se subdividem em cinco perfis de engajamento, sendo dois perfis de alto grau de engajamento no curto prazo (engajamento empenhado e espasmódico), dois perfis de alta duração do engajamento no longo prazo (engajamento

persistente e duradouro) e um perfil de comportamento intermediário (engajamento moderado);

- A maior parte dos trabalhadores que atua nos projetos apresenta um perfil de engajamento moderado;
- Em 3 dos 4 projetos estudados, trabalhadores de engajamento moderado apresentaram a maior contribuição agregada em termos do total de tempo de computação dedicado ao projeto.

Os resultados obtidos neste capítulo mostram que a análise do engajamento é satisfatória em identificar a forma como os trabalhadores atuam nos projetos. As métricas propostas ajudam a elicitar diversas características da oferta de poder cognitivo pelos trabalhadores e a analisar as semelhanças e diferenças entre eles em termos do comportamento e do total de tempo de computação agregado ao projeto.

Capítulo 4

Credibilidade de Trabalhadores em Projetos de Computação por Humanos

Mesmo que existam diversos trabalhadores engajados provendo poder cognitivo, um projeto de computação por humanos pode não ter sucesso se respostas críveis não puderem ser obtidas de tais trabalhadores. A credibilidade das respostas geradas por um ser humano em um processo de engajamento cognitivo pode ser ameaçada por uma diversidade de fatores. A rigor, não há garantia de que um ser humano gerará respostas corretas ao executar tarefas factuais e nem de que ele seguirá adequadamente as instruções fornecidas ao executar tarefas não factuais. Assim, na análise da oferta de poder cognitivo em projetos de computação por humanos, não basta uma compreensão do engajamento dos trabalhadores, é necessário também uma compreensão da credibilidade das respostas providas por eles. Nesse sentido, este capítulo descreve a pesquisa feita para ampliar o entendimento da atuação de trabalhadores em projetos de computação por humanos usando como lente o conceito de credibilidade de seres humanos.

A pesquisa aqui descrita parte da investigação de duas questões gerais sobre o que é credibilidade no contexto de tarefas de computação por humanos e como quantificar a credibilidade dos trabalhadores ao executarem esse tipo de tarefa. A partir desse entendimento, propõe-se métricas que podem ser usadas para quantificar a credibilidade de trabalhadores, e usa-se dados de projetos reais para avaliar o quão apropriadas são essas métricas e como se comportam os trabalhadores desses projetos, considerando suas credibilidades. Além disso, analisa-se em que medida os trabalhadores diferem entre si em termos dos valores de credi-

bilidade que eles apresentam ao atuarem nos projetos.

Nas seções seguintes, primeiro apresenta-se uma contextualização do conceito de credibilidade de seres humanos (Seção 4.1). Após isso, propõe-se métricas para medir a credibilidade dos trabalhadores em projetos de computação por humanos (Seção 4.2). Finalmente, as métricas propostas são utilizadas para analisar a credibilidade dos trabalhadores em dados de projetos reais. Primeiro detalham-se os materiais e métodos utilizados (Seção 4.3) e em seguida apresenta-se os resultados obtidos (Seção 4.4.

4.1 Fundamentos da Credibilidade de Seres Humanos

Busca-se nesta seção prover melhor entendimento dos conceitos e teorias tratados na literatura sobre credibilidade de seres humanos, como a definição de credibilidade, dos tipos de credibilidade, as formas de avaliação de credibilidade e os fatores que determinam a credibilidade de seres humanos em um processo de engajamento cognitivo.

4.1.1 O que é Credibilidade?

De uma forma geral, o termo credibilidade é entendido como a propriedade de ser crível, a propriedade de ser geralmente aceito, ou no qual se pode acreditar (FOGG; TSENG, 1999; WATHEN; BUREL, 2002; RIEH; DANIELSON, 2007). Ele é relacionado a diversos outros termos, como: qualidade, segurança, autoridade cognitiva e persuasão. Rieh e Danielson discutem relações entre esses termos (RIEH; DANIELSON, 2007). A qualidade de uma informação ou fonte de informação e sua credibilidade não são exatamente a mesma coisa. Credibilidade é um dos fatores que indicam qualidade, outros fatores são, por exemplo, o quão consistente é a informação e/ou o quão recente ela é. A credibilidade de uma fonte de informação está mais relacionada a ela ser crível (*believability*) do que ela ser confiável (*dependability*). Credibilidade geralmente também é associada à autoridade cognitiva. Autoridades cognitivas são fontes de informação reconhecidas como especialistas em um dado domínio. Qualidade, segurança e autoridade cognitiva são diferentes dimensões que podem ser consideradas em uma avaliação de credibilidade (FOGG; TSENG, 1999; RIEH; DANIELSON, 2007).

A noção de credibilidade de seres humanos como fontes de informação remonta ao século IV a.C., quando o filósofo grego Aristóteles conduziu seus estudos sobre lógica dedutiva e da

habilidade de oradores de convencerem uma audiência por meio de uma explanação de suas ideias (RIEH; DANIELSON, 2007). Em especial, os estudos da lógica sofista e de falácias têm grande importância no estudo de credibilidade. Por exemplo, como se verá mais adiante, no estudo de credibilidade e em estratégias utilizadas em sistemas de computação por humanos, não são incomuns abordagens como o apelo à palavra de alguma autoridade cognitiva a fim de validar um argumento (em latim *argumentum magister dixi*, apelo à palavra do mestre) e abordagens que se baseiam na ideia de que se muitas pessoas concordam com um argumento então ele é válido ou aceitável (em latim *argumentum ad populum*, apelo à multidão).

4.1.2 Tipos de Credibilidade

O julgamento que um ser humano faz em relação a acreditar ou não em uma informação provida por outro ser humano se baseia não apenas na informação propriamente dita, mas também em outros aspectos a ela associados. Diferentes tipos de julgamento podem originar diferentes tipos de credibilidade. No geral existem quatro tipos de credibilidade: presumida, reputada, aparente e experimentada (FOGG; TSENG, 1999; WATHEN; BUREL, 2002).

- Credibilidade presumida descreve o quão um ser humano acredita em uma informação ou em uma fonte de informação por causa de considerações preexistentes em sua mente. Por exemplo, seres humanos geralmente assumem que seus amigos falam a verdade sobre sua experiência com um produto e que os vendedores desses produtos mentem sobre a qualidade deles.

- Credibilidade reputada descreve o quão um ser humano acredita em uma informação ou fonte de informação porque um outro ser humano atribuiu credibilidade a ela. Por exemplo, seres humanos assumem que um produto é de qualidade se existirem boas avaliações sobre ele em relatório de avaliação feito por outros consumidores.

- Credibilidade aparente é baseada em uma avaliação superficial. Por exemplo, julgar a credibilidade dos resultados apresentados em um artigo científico considerando apenas a forma como o documento está formatado ou a qualidade da linguagem apresentada no resumo do artigo.

- Credibilidade experimentada é baseada em uma primeira experiência. Por exemplo, se

uma fonte de informação estava correta em uma primeira análise, assume-se que ela estará correta no futuro.

4.1.3 Avaliação de Credibilidade

Uma avaliação de credibilidade consiste em medir a credibilidade presumida, reputada, aparente e/ou experimentada de uma informação ou fonte de informação. A avaliação de credibilidade envolve duas fases definidas como proeminência e interpretação (FOGG, 2003). A fase de proeminência consiste em medir e colocar em perspectiva o resultado de uma avaliação de credibilidade. Por exemplo, informar que n pessoas marcaram uma determinada notícia em um *blog* como inverídica é uma forma de medir e dar perspectiva à credibilidade reputada de uma notícia. A fase de interpretação, por sua vez, é quando um ser humano julga o quanto uma informação de credibilidade afeta o uso que ele faz do item avaliado. Por exemplo, para alguns seres humanos, o fato de n pessoas marcarem uma notícia como inverídica não afeta o uso que ele faz dela, enquanto, para outros seres humanos, isso pode afetar.

Existem três modelos para avaliação de credibilidade, definidos como: avaliação binária, avaliação por limiar, avaliação espectral (FOGG; TSENG, 1999; WATHEN; BUREL, 2002). Na avaliação binária, o objeto da avaliação é percebido como crível ou não crível. Não há possibilidades intermediarias. Na avaliação por limiar existe um limite superior e inferior de credibilidade. Se a credibilidade excede o limite superior ela é considera crível. De outro modo, se a credibilidade está abaixo do limite inferior ela é considerada não crível. Se o resultado está entre os limites inferior e superior, ela é dita ser razoavelmente crível. Finalmente, na avaliação espectral não há classes de credibilidade. Toda nuance no valor de credibilidade precisa ser considerada.

O foco da pesquisa a que se refere este documento é avaliar a credibilidade dos trabalhadores e das respostas que eles provêem para as tarefas de computação por humanos que eles executam. Isso é feito considerando os diversos tipos de credibilidade em uma avaliação espectral. Ênfase é dada principalmente à fase de proeminência por meio do estudo de diferentes indicadores de credibilidade.

4.1.4 Determinantes de Credibilidade

Existem diversas teorias úteis para orientar o estudo sobre a credibilidade dos trabalhadores em sistemas de computação por humanos. Nesses sistemas, mostram-se relevantes tanto teorias estabelecidas para explicar a cognição de cada indivíduo quando teorias que envolvem a decisão coletiva de vários indivíduos. Teorias com esse propósito são normalmente referidas como *sócio-cognitivas* (STAHL, 2011). Nesse contexto a Teoria da Racionalidade Limitada (SIMON, 1972) e a Teoria do Erro Humano (REASON, 1990) são de especial interesse.

A *Teoria da Racionalidade Limitada* foi proposta por Simon (1972) com o propósito de explicar o porquê de seres humanos nem sempre atingirem uma solução ótima ou correta para os problemas aos quais são confrontados. Ele mostra que existem situações em que não é possível computar no sistema cognitivo humano todas as possíveis soluções para o problema e escolher a melhor solução. Isso implica que a racionalidade humana é limitada pela capacidade cognitiva do cérebro humano e pela complexidade do ambiente (SIMON, 1972). Nessas situações, o cérebro humano utiliza heurísticas que fazem simplificações do problema e assim são capazes de gerar soluções que o ser humano julga satisfatórias (CONLISK, 1996). No entanto, em muitos casos, essas heurísticas levam a se obter resultados incorretos ou bem distantes da solução ideal. Nesses casos, a heurística pode conter um viés cognitivo.

Exemplos de vieses cognitivos são: ancoramento (*anchoring effect*), excesso de confiança (*overconfidence*), efeito de elaboração (*framing effect*), e previsão afetiva defeituosa (*defective affective forecasting*). Efeito de ancoramento ocorre quando os seres humanos colocam demasiada importância em apenas um aspecto do problema em detrimento dos demais. Efeito de excesso de confiança é causado por uma propensão de seres humanos de superestimarem suas respostas e de atribuírem confiança excessiva para sua probabilidade de corretude. Efeito de elaboração, por sua vez, é a influência sofrida pelo ser humano pela forma como uma pergunta é elaborada. Esse efeito indica que a resposta que um ser humano provê para uma pergunta tende a ser influenciada pela forma como a pergunta é feita. Finalmente, o efeito de previsão afetiva defeituosa é a propensão de seres humanos a julgarem incorretamente a forma como alguns eventos afetariam seu estado emocional no futuro. Seres humanos tendem a colocar maior ênfase no efeito de alguns poucos fatores extremamente positivos ou extremamente negativos, ignorando diversos outros fatores relevantes.

A *Teoria do Erro Humano* foca em erros aos quais os seres humanos estão sujeitos durante a solução de um problema. Diversos fatores podem motivar erros intencionais. Já os erros não intencionais tendem a ser sistemáticos e têm relação com a forma como o sistema cognitivo humano funciona. A teoria considera que para executar uma tarefa, um ser humano passa por três estágios cognitivos: planejamento da solução, armazenamento do plano e execução do plano (REASON, 1990). Primeiro o ser humano constrói um plano mental constituído de uma sequência de passos que termina com a geração da saída para a tarefa. Como os planos não são normalmente colocados em prática imediatamente, há uma fase de armazenamento, de duração variável, que irá ocorrer entre a formulação da ação e sua execução. A fase de execução consiste na implementação do plano armazenado.

Há três níveis de controle cognitivos nesse processo: proficiência, regras e conhecimento (RASMUSSEN, 1983). O nível de controle baseado em proficiência e o nível de controle baseado em regras ocorrem em tarefas em domínios conhecidos. A distinção entre proficiência e regras está na ciência do ser humano sobre os passos seguidos na execução. Tarefas cuja execução é baseada em proficiência podem ser resolvidas pelo ser humano sem que ele seja capaz de descrever os passos que seguiu para executá-las. Este é o caso, por exemplo, de tarefas baseadas em um nível tão elevado de treinamento que a execução pelo ser humano se torna automática e inconsciente. Em tarefas cuja execução é baseada em regras, por sua vez, o ser humano sempre segue uma sequência definida de passos para executar a tarefa. Após executá-la, ele é capaz de descrever os passos seguidos. Esse é o caso, por exemplo, da solução de uma equação matemática. Por fim, o nível de controle baseado em conhecimento se refere às tarefas em domínios desconhecidos, em que o ser humano não possui experiência de solução ou regras a serem seguidas. Nesse tipo de tarefa, o ser humano constrói diferentes planos de solução e seleciona um plano considerando, por exemplo, seus objetivos pessoais e predições de efeitos dos planos.

Neste contexto, o termo *erro humano* abrange *todas as ocasiões em que uma sequência planejada de atividades mentais não consegue atingir o resultado pretendido*. Os erros assumem um número limitado de formas. Existem três formas gerais: ignorância (*mistake*), esquecimento (*lapse*) e deslize (*slip*). Como mostrado na Tabela 4.1, esses erros estão associados aos estágios cognitivos e aos níveis de controle cognitivo.

Ignorância é um erro de planejamento. Ele se manifesta na construção do plano mental

Tabela 4.1: Associações entre os tipos de erro humano, estágios cognitivo e níveis de controle cognitivo.

Tipo de Erro	Estágio Cognitivo	Nível de Controle Cognitivo
Ignorância	Planejamento	Regras e Conhecimento
Esquecimento	Armazenamento	Proficiência
Deslize	Execução	Proficiência

para resolver a tarefa. O ser humano define uma sequência de passos que ao serem seguidos não levam à resposta correta. Erros por ignorância estão relacionados ao nível de controle cognitivo baseado em regras e em conhecimento. Ignorância relacionada às regras é associada à má classificação da situação que leva à aplicação da regra errada ou à recordação incorreta de procedimentos de solução. Ignorância relacionada ao conhecimento ocorre por limitações do sistema cognitivo e conhecimento incompleto ou incorreto sobre o domínio do problema. Ou seja, o ser humano não é capaz de construir diversos planos de solução no seu sistema cognitivo e selecionar o mais adequado, ou ele não possui todo conhecimento necessário para construir um plano adequado.

Deslizes e esquecimentos, por sua vez, são erros de armazenamento e execução, geralmente classificados como erros relacionados à proficiência. Nesses casos, o ser humano possui os conhecimentos necessários para executar a tarefa e ele cria um plano mental com a sequência de passos que precisam ser seguidos para se chegar à resposta correta. No entanto, no caso do esquecimento, ocorre um erro na fase de armazenamento. O ser humano simplesmente se esquece de executar algum passo. No caso do deslize, por sua vez, há um erro na fase de execução. O ser humano executa incorretamente algum dos passos planejados.

Os problemas associados à racionalidade limitada, aos vieses cognitivos e aos tipos de erros humanos são alguns dos possíveis fatores que podem levar trabalhadores a não gerarem respostas críveis para as tarefas que eles executarem. Os diversos aspectos humanos descritos na Figura 2.3 individualmente ou em conjunto também podem exercer algum efeito (PONCIANO et al., 2014). A maior ênfase neste trabalho está em caracterizar a credibilidade dos trabalhadores. Nesse contexto, é relevante distinguir situações em que cada trabalhador sistematicamente apresenta respostas que não são críveis (que pode ocorrer em razão de vieses cognitivos ou ignorância) e situações em que isso não ocorre com frequência (que pode indicar deslizes e esquecimento).

4.2 Medindo Credibilidade em Computação por Humanos

Nesta seção são propostas métricas para medir a credibilidade de trabalhadores em projetos de computação por humanos. Considera-se que a credibilidade de cada trabalhador pode variar com o grau de dificuldade da tarefa. Dado que a dificuldade afeta as respostas geradas pelos seres humanos, quanto maior for a divergência nas respostas para uma tarefa, mais difícil a tarefa tende a ser para o grupo de trabalhadores que a está executando (ARCANJO et al., 2014; AROYO; WELTY, 2014). O grau de dificuldade de uma tarefa é medido usando entropia de Shannon (SHANNON, 1951). Essa entropia mede o grau de divergência existente no conjunto de respostas providas por diferentes trabalhadores. O grau de dificuldade é denotado por h e definido pela Equação 4.1.

$$h = -\sum_{g \in G} \Big(\Pr(g) \times \log_2 \Pr(g) \Big) \tag{4.1}$$

Nessa equação, G denota o conjunto de respostas obtidas para a tarefa, cada $g \in G$ denota uma resposta distinta, e $\Pr(g)$ denota a proporção de trabalhadores que proveram a resposta g. Quando todos os trabalhadores proveram uma mesma resposta para a tarefa, o grau de dificuldade assume o valor mínimo ($h = 0$). O valor de h cresce à medida que aumenta a diversidade das respostas recebidas para a tarefa e que os trabalhadores se dividem igualdade entre essas respostas. Para cada trabalhador, estima-se um valor de credibilidade para cada grau de dificuldade de tarefa que ele executada. Assim, as tarefas executadas pelos trabalhadores são categorizadas pelo seu grau de dificuldade. Essa categorização é feita arredondando os valores de h para uma casa decimal, ou seja, $h = 0, 0.1, 0.2, \ldots$.

Em tarefas de computação por humanos, uma resposta de escolha consensual ou de escolha da maioria dos trabalhadores é uma resposta geralmente aceita pelos usuários (SHESHADRI; LEASE, 2013; RAO; HUANG; FU, 2013; AROYO; WELTY, 2014; PONCIANO et al., 2014). Um trabalhador que gera respostas em discordância com a maioria sistematicamente é um trabalhador com baixa credibilidade. Dessa forma, as métricas de credibilidade propostas são baseadas no conceito de concordância com a maioria. Um trabalhador é considerado tanto crível quanto for a probabilidade da resposta provida por ele ser igual à resposta que seria provida pela maioria dos trabalhadores que atuam no projeto. A partir dessa ideia, são definidas quatro métricas de credibilidade dos trabalhadores: concordância simples, concor-

dância experimentada, concordância ponderada e concordância reputada. Como descrito nos próximos parágrafos, essas métricas envolvem as diferentes formas de avaliar credibilidade.

Concordância simples. Trata-se de uma métrica de credibilidade superficial definida como a proporção de tarefas em que o trabalhador proveu uma resposta igual à resposta dada pela maioria dos trabalhadores que executaram a tarefa. Seja $n_{w,h}$ o número total de tarefas com grau de dificuldade h executada por um trabalhador w, e seja $f_{w,h}$ o montante dessas tarefas em que o trabalhador w proveu a mesma resposta provida pela maioria dos trabalhadores que executaram a tarefa. A probabilidade de concordância entre as respostas fornecidas pelo trabalhador w e as respostas providas pela maioria dos trabalhadores é calculada conforme definido pela Equação 4.2. Por essa equação, quando $c_{w,h} = 1$, há uma concordância completa entre a resposta fornecida pelo trabalhador w e a resposta dada pela maioria em todas as tarefas executadas pelo trabalhador w. De modo contrário, quando $c_{w,h}$ se aproxima de 0, menor tende a ser a concordância entre o trabalhador w e a maioria.

$$c_{w,h} = \frac{f_{w,h}}{n_{w,h}}, c_{w,h} \in [0, 1] \tag{4.2}$$

Concordância experimentada. Esta é uma métrica de credibilidade experimentada baseada na estatística Cohen's kappa (COHEN, 1960). Essa estatística tem sido usada para medir a concordância entre as respostas providas por duas pessoas. Aqui, essa estatística é usada para medir o grau de concordância entre as respostas providas por um trabalhador e as respostas providas pela maioria dos trabalhadores para as mesmas tarefas. Ao contrário do que ocorre na concordância superficial, a concordância experimentada leva em conta não apenas probabilidade de concordância ($c_{w,h}$). Ela também considera a probabilidade de concordância que pode ocorrer por acaso ($z_{w,h}$). Ou seja, a probabilidade de os trabalhadores concordarem dando respostas aleatórias para as tarefas. A concordância experimentada é denotada por $e_{w,h}$ e formalizada na Equação 4.3. Para obter os valores no intervalo entre 0 e 1, como nas outras métricas, aplica-se o cálculo $(e_{i,h} + 1)/2$. Quando o resultado é 1, há plena concordância entre o trabalhador w e a maioria. Se ele é maior que 0,5, o grau de concordância é maior ou igual ao que se ocorreria apenas em razão do acaso. Finalmente, um resultado menor que 0,5 indica que o grau de concordância é inferior ao que ocorreria apenas em razão

do acaso.

$$e_{i,h} = \frac{c_{w,h} - z_{w,h}}{1 - z_{w,h}}, e_{w,h} \in [-1, 1] \tag{4.3}$$

Concordância ponderada. Esta é uma métrica de credibilidade presumida baseada na ideia de quanto mais informação foi utilizada para estimar a credibilidade, mais provável é que a estimativa seja correta. Por exemplo, a estimativa de credibilidade com base em apenas uma resposta dada por um trabalhador parece ser menos confiável do que a estimativa de credibilidade com base em 10 respostas fornecidas por um trabalhador. A métrica de concordância ponderada implementa essa ideia. Ela é denotada por $p_{w,h}$ e definida na Equação 4.4. Trata-se de uma média harmônica ponderada entre a proporção de concordância ($c_{w,h}$) e a concordância neutra de 0,5. O peso de $c_{w,h}$ é o número de tarefas executadas pelo trabalhador ($n_{w,h}$) e o peso de 0,5 é de 1. Dessa forma, quando $n_{w,h} = 0$, a concordância ponderada $p_{w,h}$ é de 0,5 e, na medida em que $n_{w,h}$ aumenta, o valor de $p_{w,h}$ tende a $c_{w,h}$.

$$p_{w,h} = \frac{n_{w,h} + 1}{\frac{n_{w,h}}{c_{w,h}} + \frac{1}{0.5}}, p_{w,h} \in [0, 1] \tag{4.4}$$

Concordância reputada. Esta é um métrica de credibilidade reputada. A ideia implementada nesta métrica é de que a credibilidade de um trabalhador deve ser aumentada quando ele concorda com trabalhadores altamente credíveis e deve ser diminuída quando ele discorda de trabalhadores altamente credíveis. Assim, o propósito da métrica de concordância reputada é considerar na credibilidade de um trabalhador w escores de credibilidade dos outros trabalhadores com quem ele concordou e discordou no passado. Seja $K_{w,h}$ o conjunto de trabalhadores com os quais w concordou quando ele proveu uma resposta igual à resposta da maioria. A credibilidade média desse conjunto de trabalhadores é dada por $k_{w,h}$ e calculada como na Equação 4.5. Seja $M_{w,h}$ o conjunto de trabalhadores que proveram a resposta da maioria quando o trabalhador w não gerou uma resposta igual à resposta provida pela maioria. A credibilidade média desse conjunto de trabalhadores é dada por $m_{w,h}$ e calculada como na Equação 4.6. Uma vez definidos $k_{w,h}$ e $m_{w,h}$, pode-se calcular a concordância reputada do trabalhador w como na Equação 4.7.

$$k_{w,h} = \frac{\sum_{i \in K_{w,h}} c_{i,h}}{|K_{w,h}|} \tag{4.5}$$

$$m_{w,h} = \frac{\sum_{i \in M_{w,h}} c_{i,h}}{|M_{w,h}|} \tag{4.6}$$

$$r_{w,h} = \frac{c_{w,h} + k_{w,h} - m_{w,h} + 1}{3}, r_{w,h} \in [0, 1] \tag{4.7}$$

A credibilidade $r_{w,h}$ assume o valor mínimo 0 quando ocorrem as seguintes condições: (i) o trabalhador w discordou da maioria em todas as tarefas que ele executou (i.e., $c_{w,h} = 0$), (ii) dado que o trabalhador w nunca concordou com a maioria, então não há ganho de credibilidade de outros trabalhadores que forneceram a resposta da maioria (i.e., $K_{w,h}$ é um conjunto vazio e portanto $k_{w,h} = 0$), e (iii) os trabalhadores que geraram a resposta da maioria, quando w não forneceu a resposta majoritária, têm a mais alta credibilidade, de modo que o trabalhador w perde 1 em credibilidade por discordar de tais trabalhadores (i.e., $m_{w,h} = 1$). De outro modo, a credibilidade $r_{w,h}$ assume o valor máximo 1 quando as seguintes condições são satisfeitas: (i) o trabalhador w concordou com a maioria em todas as tarefas que ele executou (i.e., $c_{w,h} = 1$), (ii) os trabalhadores com quem ele concordou tem a mais alta credibilidade, o que faz com que ele ganhe o máximo de credibilidade destes trabalhadores (i.e., $k_{w,h} = 1$), e (iii) o trabalhador w nunca discordou de trabalhadores que prestavam a resposta da maioria, então, não há perda de credibilidade por não concordar com esses trabalhadores (i.e., $M_{w,h}$ é um conjunto vazio e, portanto, $m_{w,h} = 0$).

Essas quatro métricas de credibilidade baseadas em concordância se encaixam no objetivo deste trabalho de considerar uma diversidade dos aspectos de credibilidade dos trabalhadores. Concordância superficial é a métrica de credibilidade mais simples; ela leva em conta apenas o grau de concordância entres os trabalhadores. Concordância experimentada, por sua vez, mede o grau de concordância real ao levar em conta a concordância que seria esperada e deduzindo-se a quantidade de concordância que pode ocorrer devido ao acaso. Concordância presumida, por sua vez, pondera o grau de concordância com a quantidade de dados utilizados no seu cálculo. Finalmente, a métrica concordância reputada leva em conta não apenas a quantidade de concordância exibida por um trabalhador no passado, mas

também a credibilidade dos outros trabalhadores com os quais ele concordou e discordou.

Todas as métricas se valem de conceitos da literatura de avaliação de credibilidade e de uma perspectiva sócio-cognitiva do processo de execução de tarefas pelos trabalhadores. As métricas em si são de avaliação espectral. Ou seja, elas não definem categorias como "crível" e "não crível". O resultado da avaliação de credibilidade é um número decimal que indica o quão crível o trabalhador é ao executar tarefas no projeto. Nesse sentido, elas visam o objetivo de proeminência da avaliação de credibilidade. A interpretação do valor de credibilidade, por sua vez, pode depender da análise do usuário. Nesse sentido, o usuário pode definir o limiar de credibilidade requerida para que a resposta provida por um trabalhador seja considerada crível.

4.3 Materiais e Métodos de Avaliação

A análise das métricas de credibilidade é realizada usando bases de dados obtidas de projetos reais de computação por humanos. Tais bases de dados são descritas na próxima seção. Em seguida, os métodos utilizados na avaliação são detalhados.

4.3.1 Descrição dos Projetos Estudados

O estudo da credibilidade dos trabalhadores em sistemas de computação por humanos impõe diversos requisitos em termos de base de dados. Para cada evento de execução de tarefa, a base de dados deve conter a informação do trabalhador que fez a execução, a tarefa a que se refere a execução e a resposta provida pelo trabalhador. Além disso, é importante que exista redundância na execução das tarefas. Ou seja, diversos trabalhadores devem ter provido respostas para uma mesma tarefa. Isso é necessário para que se possa calcular a dificuldade de cada tarefa e estimar a credibilidade de cada trabalhador.

Foram obtidas cinco bases de dados com essas características: Análise de Sentimentos, Sun4All, Cell Spotting e Julgamento de Fatos[1]. Nos próximos parágrafos são descritos os dados disponíveis nessas bases de dados que são relevantes ao estudo da credibilidade dos

[1]Os dados desse projeto foram disponibilizados pela empresa Google. Para mais informações sobre essa disponibilização, consulte a publicação de lançamento na URL http://googleresearch.blogspot.com.br/2013/04/50000-lessons-on-how-to-read-relation.html.

trabalhadores. As bases de dados dos projetos Análise de Sentimentos, Sun4All e Cell Spotting foram utilizadas no estudo do engajamento dos trabalhadores apresentado no capítulo anterior. Essas bases de dados são descritas novamente no próximo parágrafo destacando as informações relevantes ao estudo de credibilidade.

A base de dados Análise de Sentimentos consiste em tarefas de julgamentos da condição climática relatada em *tweets*. Trata-se de 569.375 eventos de execuções de tarefas gerados por 1.960 trabalhadores. Existe um total de 98.980 tarefas e respostas de pelo menos 5 trabalhadores por tarefa. O projeto Sun4All consiste em tarefas de contagem de manchas solares. Trata-se de 4.328 eventos de execuções de tarefas gerados por 116 trabalhadores diferentes. Existe um total de 417 tarefas e respostas de pelo menos 8 trabalhadores por tarefa. O projeto Cell Spotting consiste em tarefas contagem de células mortas. Trata-se de 94.137 eventos de execuções de tarefas gerados por 1.103 trabalhadores. Existe um total de 4.067 tarefas e respostas de pelo menos 8 trabalhadores por tarefa.

A base de dados Julgamento de Fatos consiste em tarefas de julgamento de relações referentes a pessoas públicas com base em dados existentes na Wikipédia. Cada tarefa apresenta ao trabalhador uma relação do tipo "a pessoa X se graduou na universidade Y". Os trabalhadores foram solicitados a julgar se a relação é "verdadeira", "falsa" ou "não responder". Existem dados de 220.000 eventos de execuções de tarefas gerados por 57 trabalhadores diferentes. Existe um total de 42.624 tarefas e respostas de pelos menos 5 trabalhadores por tarefa.

Esses projetos são compostos por tarefas com diferentes graus de dificuldades. Existem 6 diferentes graus de dificuldade de tarefas no projeto Julgamento de Fatos, 24 diferentes graus de dificuldade de tarefas no projeto Análise de Sentimentos, 28 diferentes graus de dificuldade no projeto Sun4All e 35 diferentes graus de dificuldade de tarefas no projeto Cell Spotting. A distribuição da dificuldade das tarefas nos projetos varia de um projeto para outro (Fig. 4.1). Por exemplo, no projeto Julgamento de Fatos 70% das tarefas são fáceis (dificuldade 0), há unanimidade nas respostas providas. Por outro lado, no projeto Cell Spotting, tarefas fáceis consiste em menos de 1%.

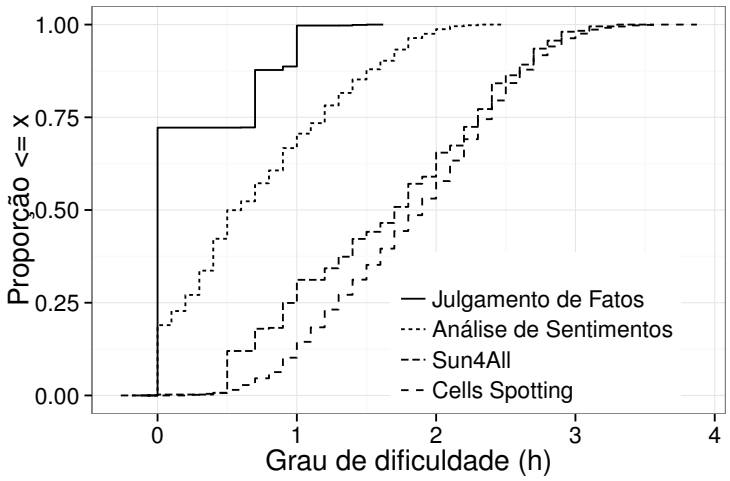


Figura 4.1: Distribuição do grau de dificuldade das tarefas nos projetos Julgamento de Fatos, Análise de Sentimentos, Sun4All e Cell Spotting.

4.3.2 Método de Caracterização de Semelhanças e Diferenças entre Trabalhadores

A análise das semelhanças e diferenças entre os trabalhadores em um dado projeto é realizada considerando as FDAs das métricas de credibilidade. Essas distribuições informam o quão os trabalhadores se concentram em uma determinada faixa de valores para uma dada métrica em análise. Em algumas situações, mostra-se importante investigar em que medida duas FDAs são semelhantes. Por exemplo, verificar se a distribuição dos trabalhadores pela métrica de credibilidade concordância simples é igual à distribuição dos trabalhadores pela métrica de credibilidade concordância reputada. Outro caso de interesse é aquele em que se deseja saber se a distribuição dos trabalhadores em uma métrica em um dado projeto é semelhante à distribuição dos trabalhadores pela mesma métrica em outro projeto. Essas comparações de FDAs são realizadas por meio da estatística D do teste Two-sample Kolmogorov-Smirnov (SMIRNOV, 1939). Essa estatística indica a distância entre duas FDAs. Na análise dos resultados, para dizer que duas distribuições são iguais, considera-se o grau de significância de $0,05$.

4.3.3 Método de Caracterização e Análise de Relações entre Métricas de Credibilidade

As relações entre diferentes métricas de credibilidade são analisadas tendo em vista responder duas questões principais: *(i)* quão distantes são os valores de credibilidade estimados por diferentes métricas; e *(ii)* em que medida métricas diferentes geram ranques iguais de trabalhadores.

A distância entre os valores de credibilidade estimados por diferentes métricas é medida usando a *distância absoluta média*. Seja x e y duas métricas de credibilidade, a distância absoluta média entre os valores de credibilidade estimados por essas métricas para os trabalhadores no conjunto W é definida como na Equação 4.8. Essa distância assume o valor 0 quando os valores de credibilidade estimados pelas métricas forem todos iguais. Essa distância, por outro lado, assume o valor 1 quando uma métrica estima valores de credibilidade iguais a 0 para todos os trabalhadores enquanto uma outra métrica estima os valores de credibilidade iguais a 1 para todos os trabalhadores.

$$d(x, y) = \frac{1}{|W|} \times \sum_{w \in W} |x_w - y_w| \tag{4.8}$$

A distância entre ranques de credibilidade gerados por diferentes métricas é medida usando a *distância de Kendall (τ)* (FAGIN; KUMAR; SIVAKUMAR, 2003). Essa distância mede a proporção de mudanças que precisariam ser feitas em um dos ranques para que ele se torne igual ao outro. A distância assume o valor 0 quando ranques gerados por diferentes métricas são iguais. Por outro lado, a distância assume o valor 1 quando esses ranques estiverem em ordem inversa.

As análises da distância absoluta média e da distância de Kendall são complementares. Observe que mesmo métricas de credibilidade que estimam valores de credibilidade distantes entre si podem originar ranques de credibilidade iguais.

4.4 Apresentação e Análise dos Resultados

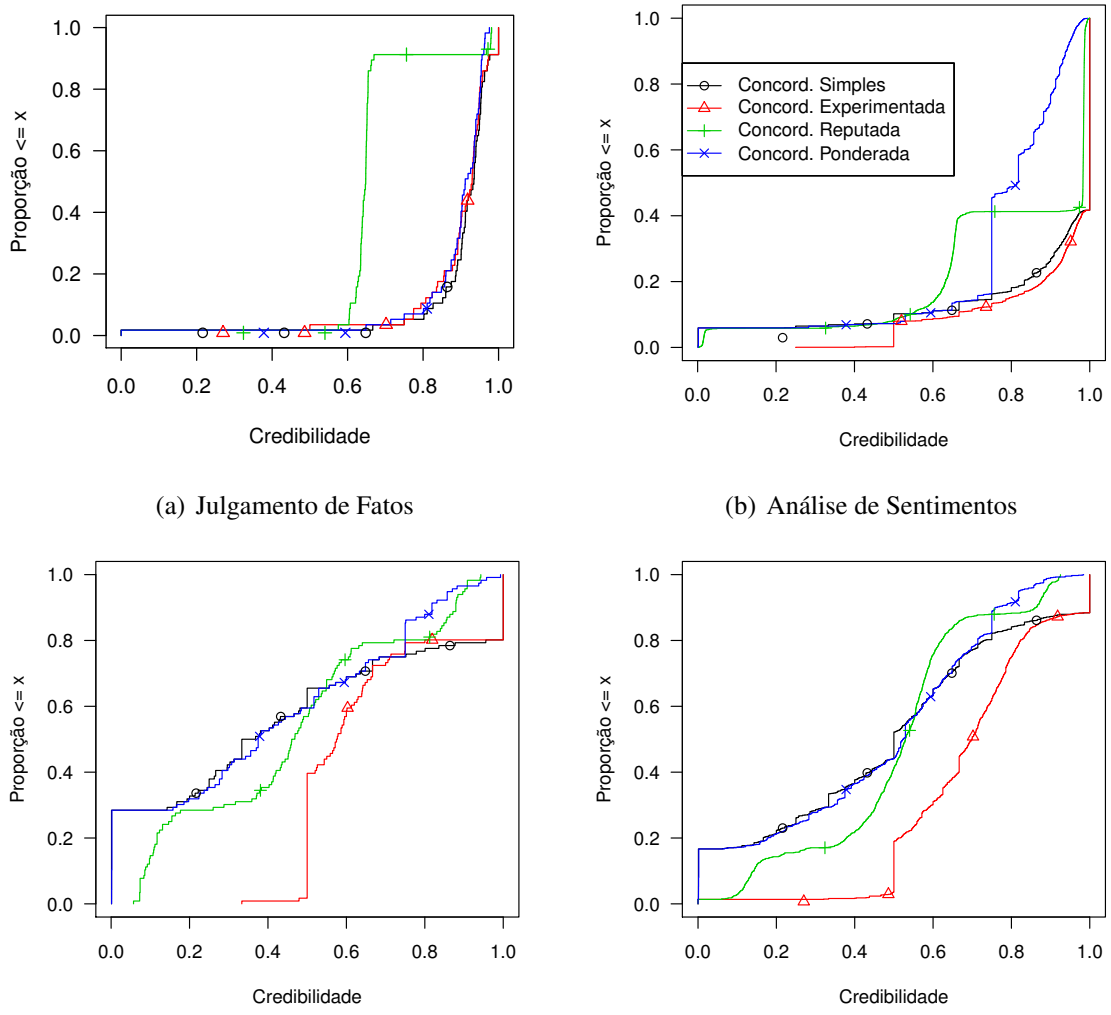
Nesta seção são discutidos os resultados obtidos ao utilizar as métricas de credibilidade propostas. Primeiro, discutem-se as distribuições dos trabalhadores em termos de credibilidade. Em seguida, analisa-se como a credibilidade dos trabalhadores varia quando diferentes métricas são utilizadas.

4.4.1 Distribuições de Credibilidade

A Figura 4.1 mostra a distribuição dos valores de credibilidade dos trabalhadores nos projetos Julgamento de Fatos (Fig 4.2(a)), Análise de Sentimentos (Fig 4.2(b)), Sun4All (Fig 4.2(c)) e Cell Spotting (Fig 4.2(d)). Os valores de credibilidade exibidos são calculados usando todas as tarefas juntas, sem distinção do grau de dificuldade. Nota-se que as distribuições têm formatos diferentes dependendo do projeto estudado e da métrica de credibilidade utilizada. Essas diferenças são discutidas nos próximos parágrafos.

No projeto Julgamento de Fatos (Fig 4.2(a)) as distribuições dos valores de credibilidades estimados pelas métricas concordância simples, concordância experimentada e concordância ponderada são, no geral, bastante próximas. Isso é uma característica esperada nesse projeto em que no geral os trabalhadores são bastante qualificados e apresentam unanimidade de resposta em 70% das tarefas. Já os valores de credibilidade estimados pela métrica de concordância reputada apresentam características diferenciadas. Por essa métrica, grande parte dos trabalhadores tende a exibir valores de credibilidade menores comparado às demais métricas. Por exemplo, enquanto 90% dos trabalhadores exibem concordância reputada menor ou igual a 0,65, mais de 90% dos trabalhadores exibem credibilidade superior a esse valor quando as demais métricas são utilizadas. Isso ocorre por que a concordância reputada é uma métrica de credibilidade mais conservadora. Por ela, os trabalhadores não recebem um alto incremento de credibilidade quando concordam com trabalhadores que não são muito críveis. Eles podem, inclusive, sofrer um decremento de credibilidade quando discordam de outros trabalhadores que exibem valores de credibilidade elevados. Isso é especialmente importante neste projeto que possui poucos trabalhadores e em que todos executaram muitas tarefas.

No projeto Análise de Sentimentos (Fig. 4.2(b)), pode-se observar que a métrica de con-



(a) Julgamento de Fatos

(b) Análise de Sentimentos

(c) Sun4All

(d) Cell Spotting

Figura 4.2: Distribuição dos valores de credibilidade dos trabalhadores medida pelas métricas concordância simples, concordância experimentada, concordância reputada e concordância ponderada nos projetos Julgamento de Fatos, Análise de Sentimentos, Sun4All e Cell Spotting.

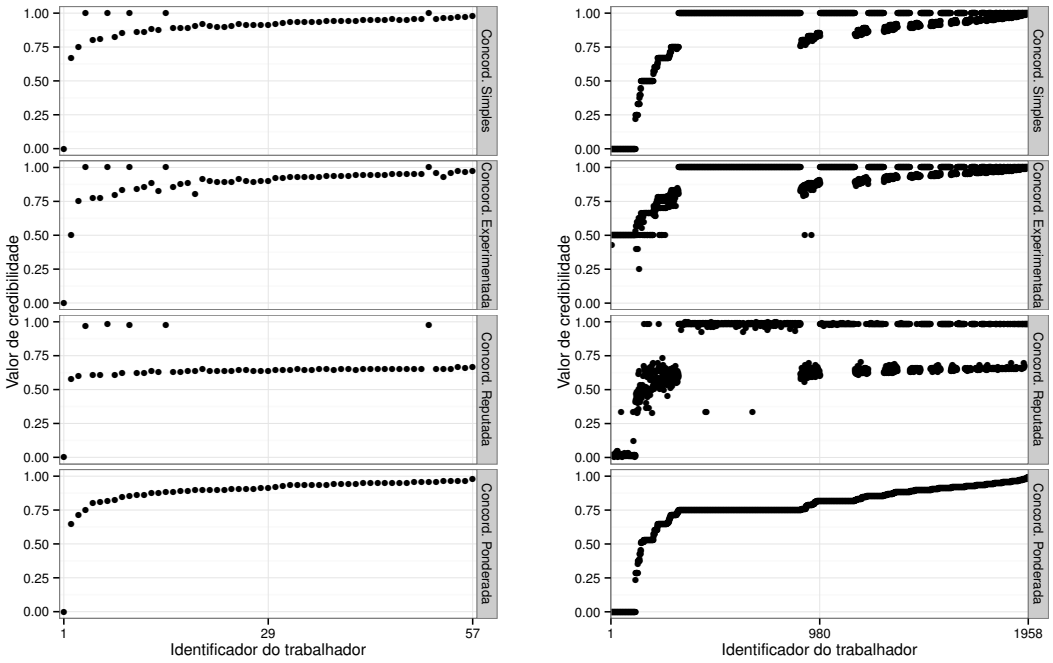
cordância ponderada revela sua principal característica que é ser conservadora na estimativa da credibilidade de trabalhadores que executaram poucas tarefas. Nesse projeto também se observa, porém, com menor intensidade, a característica da métrica de concordância reputada destacada na análise dos resultados obtidos no projeto de Julgamento de Fatos.

Os projetos Sun4All (Fig. 4.2(c)) e Cell Spotting (Fig. 4.2(c)) apresentam características gerais bastante semelhantes. Um ponto a ser destacado é que nesses projetos a métrica concordância experimentada é a menos conservadora. Ou seja, ela gera valores de credibilidade geralmente maiores que as demais, principalmente em relação às métricas de concordância superficial e presumida. Isso ocorre quando a probabilidade dos trabalhadores concordarem de forma aleatória é baixa. Essa é uma característica desses projetos que são compostos de tarefas em que não há um conjunto predefinido de respostas, como ocorre nos demais projetos.

4.4.2 Credibilidade dos Trabalhadores em Diferentes Métricas

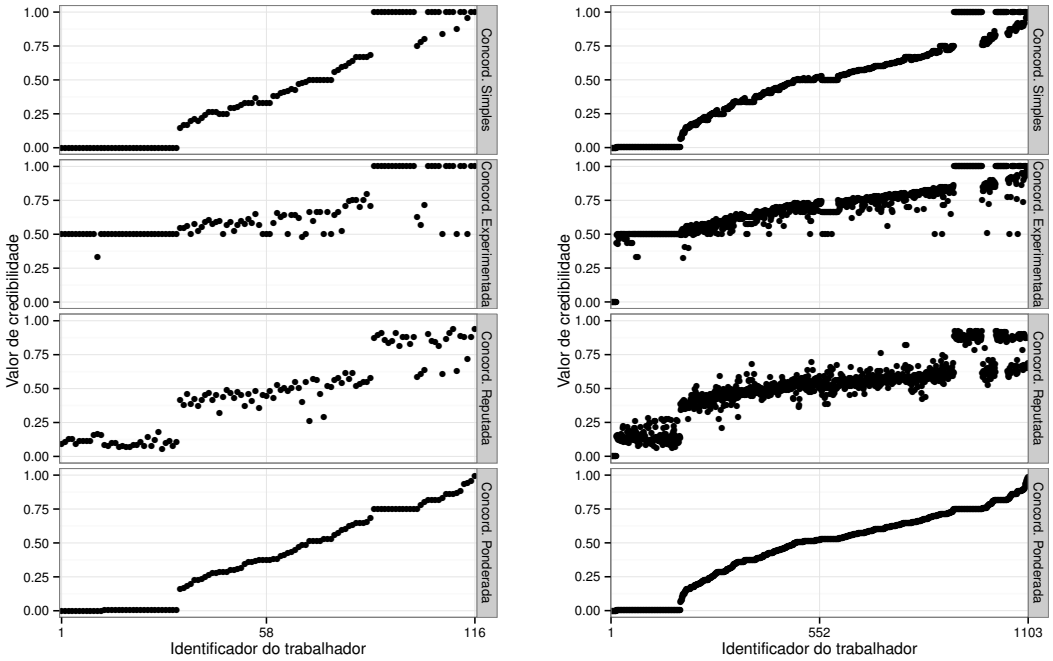
A Figura 4.3 mostra os valores de credibilidade dos trabalhadores estimados pelas métricas de credibilidade. Os trabalhadores encontram-se ranqueados em ordem crescente pelos valores de credibilidade que eles apresentam na métrica concordância ponderada. Dois comportamentos são nítidos nessa figura: os valores de credibilidade estimados por uma métrica não são iguais aos valores de credibilidade estimados por outra métrica e o ranque de credibilidade dos trabalhadores gerado usando os valores estimados por uma métrica não é igual ao ranque de credibilidade gerados usando os valores estimados pelas outras métricas.

Quando se considera que os trabalhadores podem exibir um valor de credibilidade para cada grau de dificuldade de tarefa que eles executam, pode-se obter ranques de credibilidade dos trabalhadores para cada grau de dificuldade de tarefas. Nesse contexto, mostra-se relevante entender em que medida os valores de credibilidade e os ranques estimados por diferentes métricas diferem entre si. Considerando os 5 graus de dificuldade de tarefas que envolvem as maiores quantidades de trabalhadores, a Tabela 4.2 mostra a distância absoluta média entre os valores de credibilidade estimados pelas métricas de credibilidade em cada grau de dificuldade de tarefas e a Tabela 4.3 mostra a distância de Kendall (τ) entre pares de ranques de trabalhadores gerados usando as métricas de credibilidade em cada grau de dificuldade.



(a) Julgamento de Fatos

(b) Análise de Sentimentos



(c) Sun4All

(d) Cell Spotting

Figura 4.3: Credibilidade dos trabalhadores medida pelas métricas concordância simples, concordância experimentada, concordância reputada e concordância ponderada nos projetos: Julgamento de Fatos, Análise de Sentimentos, Sun4All e Cell Spotting. Os trabalhadores estão ranqueados pelos valores que apresentam na métrica concordância ponderada. Tarefas de todos os graus de dificuldade foram incluídas no cálculo da credibilidade.

Tabela 4.2: Distância absoluta média entre os valores de credibilidade estimados pelas métricas concordância simples (c), concordância experimentada (e), concordância reputada (r) e concordância ponderada (v) em cada grau de dificuldade de tarefas.

Julgamento de Fatos

Par	**h=0,0** $N = 39$	**h=0,7** $N = 39$	**h=0,9** $N = 39$	**h=1** $N = 39$	**h=1.4** $N = 39$
$d(c, e)$	**0,00**	0,04	0,05	0,02	0,11
$d(c, r)$	**0,00**	0,22	0,13	0,11	**0,05**
$d(c, p)$	**0,00**	**0,00**	**0,03**	**0,00**	0,09
$d(e, r)$	**0,00**	0,20	0,14	0,10	0,14
$d(e, p)$	**0,00**	0,03	0,06	0,02	0,19
$d(r, p)$	**0,00**	0,22	0,11	0,11	0,08

Análise de Sentimentos

Par	**h=0,0** $N = 465$	**h=0.1** $N = 465$	**h=0.2** $N = 465$	**h=0.3** $N = 465$	**h=0.4** $N = 465$
$d(c, e)$	**0,00**	**0,00**	**0,00**	**0,0**1	**0,02**
$d(c, r)$	**0,00**	0,01	0,02	0,05	0,05
$d(c, p)$	0,13	0,12	0,11	0,09	0,11
$d(e, r)$	**0,00**	0,02	0,03	0,06	0,06
$d(e, p)$	0,13	0,12	0,12	0,10	0,13
$d(r, p)$	0,13	0,12	0,12	0,12	0,13

Sun4All

Par	**h=0,5** $N = 30$	**h=0,7** $N = 28$	**h=0,9** $N = 30$	**h=1,8** $N = 28$	**h=2,2** $N = 30$
$d(c, e)$	0,32	**0,03**	0,09	0,21	0,32
$d(c, r)$	0,13	0,08	0,11	0,10	0,12
$d(c, p)$	**0,04**	0,09	**0,07**	**0,04**	**0,03**
$d(e, r)$	0,25	0,07	0,14	0,20	0,25
$d(e, p)$	0,34	0,11	0,15	0,24	0,34
$d(r, p)$	0,12	0,12	0,11	0,10	0,12

Cell Spotting

Par	**h=1,7** $N = 234$	**h=2,1** $N = 234$	**h=2,2** $N = 234$	**h=2,3** $N = 234$	**h=2,4** $N = 234$
$d(c, e)$	0,18	0,25	0,25	0,28	0,29
$d(c, r)$	0,11	0,11	0,10	0,11	0,12
$d(c, p)$	**0,04**	**0,02**	**0,02**	**0,02**	**0,02**
$d(e, r)$	0,21	0,20	0,21	0,21	0,20
$d(e, p)$	0,21	0,25	0,26	0,26	0,29
$d(r, p)$	0,10	0,10	0,09	0,09	0,11

Nota 1: Encontram-se em negrito os menores valores de distância em cada grau de dificuldade de tarefas.
Nota 2: São exibidos resultados para os 5 graus de dificuldade de tarefas que envolvem as maiores quantidades de trabalhadores (N).

Tabela 4.3: Distância de Kendall (τ) entre pares de ranques de trabalhadores gerados usando as métricas de credibilidade concordância simples (c), concordância experimentada (e), concordância reputada (r) e concordância ponderada (v) em cada grau de dificuldade de tarefas.

Julgamento de Fatos

Par	**h=0,0** $N = 39$	**h=0,7** $N = 39$	**h=0,9** $N = 39$	**h=1** $N = 39$	**h=1.4** $N = 39$
$\tau(c, e)$	**0,00**	0,38	0,44	0,45	**0,30**
$\tau(c, r)$	**0,00**	0,48	0,45	0,42	0,40
$\tau(c, p)$	0,52	**0,34**	0,34	**0,04**	0,34
$\tau(e, r)$	**0,00**	0,52	**0,33**	0,45	0,36
$\tau(e, p)$	0,52	0,41	0,60	0,47	0,39
$\tau(r, p)$	0,52	0,36	0,52	0,43	0,40

Análise de Sentimentos

Par	**h=0,0** $N = 465$	**h=0,1** $N = 465$	**h=0,2** $N = 465$	**h=0,3** $N = 465$	**h=0,4** $N = 465$
$\tau(c, e)$	**0,00**	**0,04**	**0,08**	**0,17**	**0,17**
$\tau(c, r)$	**0,00**	0,49	0,50	0,50	0,50
$\tau(c, p)$	0,45	0,46	0,47	0,46	0,47
$\tau(e, r)$	**0,00**	0,49	0,50	0,50	0,51
$\tau(e, p)$	0,45	0,46	0,47	0,47	0,47
$\tau(r, p)$	0,45	0,47	0,47	0,49	0,47

Sun4All

Par	**h=0,5** $N = 30$	**h=0,7** $N = 28$	**h=0,9** $N = 30$	**h=1,8** $N = 28$	**h=2,2** $N = 30$
$\tau(c, e)$	**0,14**	**0,25**	0,29	0,37	**0,23**
$\tau(c, r)$	0,50	0,33	0,53	0,54	0,41
$\tau(c, p)$	0,51	0,46	0,47	**0,3**2	0,30
$\tau(e, r)$	0,50	0,38	0,54	0,59	0,44
$\tau(e, p)$	0,52	0,48	0,49	0,48	0,36
$\tau(r, p)$	0,39	0,42	**0,28**	**0,32**	0,40

Cell Spotting

Par	**h=1,7** $N = 234$	**h=2,1** $N = 234$	**h=2,2** $N = 234$	**h=2,3** $N = 234$	**h=2,4** $N = 234$
$\tau(c, e)$	0,47	0,48	0,50	0,44	0,50
$\tau(c, r)$	0,51	0,49	0,49	0,50	0,50
$\tau(c, p)$	**0,40**	**0,38**	**0,38**	**0,38**	**0,41**
$\tau(e, r)$	0,52	0,50	0,47	0,53	0,48
$\tau(e, p)$	0,48	0,50	0,51	0,50	0,45
$\tau(r, p)$	0,47	0,50	0,51	0,47	0,50

Nota 1: Encontram-se em negrito os menores valores de distância em cada grau de dificuldade de tarefas.
Nota 2: São exibidos resultados para os 5 graus de dificuldade de tarefas que envolvem as maiores quantidades de trabalhadores (N).

Os resultados apresentados na Tabela 4.2 indicam que *métricas diferentes geralmente estimam diferentes valores de credibilidade*. A distância média entre os valores estimados por diferentes métricas tende a variar com o projeto e com a dificuldade das tarefas. Nos projetos Julgamento de Fatos e Análise de Sentimentos as métricas tendem a estimar valores de credibilidade muito próximos (uma diferença entre eles de 0 a 0,13) em tarefas com grau de dificuldade 0. Nos projetos Sun4All e CellSpotting, as distâncias são sempre maiores que 0. Nesses projetos, no geral, as métricas que estimam valores de credibilidade mais próximos são as métricas concordância simples e concordância ponderada.

Os resultados apresentados na Tabela 4.3, por sua vez, indicam que, *quando utilizadas para gerar ranques dos trabalhadores, as métricas propostas geralmente geram ranques diferentes entre si*. Ou seja, a ordem de credibilidade dos trabalhadores muda dependendo da métrica de credibilidade utilizada para ordená-los. Nos projetos Julgamento de Fatos e Análise de Sentimentos, os ranques serão iguais em tarefas com dificuldade 0, exceto quando a métrica concordância ponderada é utilizada. Já nos projetos, Sun4All e Cell Spotting os ranques são sempre diferentes independentemente das métricas de credibilidade utilizadas e do grau de dificuldade das tarefas.

4.5 Considerações Finais

Neste capítulo, discutiu-se o uso do conceito de credibilidade de seres humanos na caracterização da oferta de poder cognitivo pelos trabalhadores em projetos de computação por humanos. Têm-se duas principais contribuições neste estudo. A primeira contribuição é a contextualização do conceito de credibilidade dos trabalhadores ao atuarem em projetos de computação por humanos. A segunda contribuição é a proposta de quatro métricas para analisar a credibilidade de trabalhadores em um projeto de computação por humanos considerando diversos comportamentos de concordância entre eles. Tais métricas são definidas como: concordância simples, concordância experimentada, concordância ponderada e concordância reputada.

As métricas propostas foram utilizadas na análise da oferta de poder cognitivo pelos trabalhadores em quatro projetos reais de computação por humanos. Os projetos diferem entre si em termos do tipo de tarefa desempenhada pelo trabalhador, da quantidade total de

trabalhadores, do número de tarefas executadas por eles, da quantidade de trabalhadores que atuaram em cada tarefa e da diversidade de dificuldade de tarefas. Os principais resultados obtidos nos projetos analisados foram:

- Métricas diferentes geralmente estimam diferentes valores de credibilidade. A distância média entre os valores estimados por diferentes métricas tende a variar com o projeto e com a dificuldade das tarefas;
- As métricas concordância reputada e concordância ponderada tendem a ser mais conservadoras gerando valores de credibilidade geralmente menores e as métricas de concordância simples e experimentada tendem a ser menos conservadoras gerando valores de credibilidade maiores;
- A ordem de credibilidade dos trabalhadores que atuam no projeto muda dependendo da métrica de credibilidade utilizada. Ou seja, trabalhadores apontados como os mais críveis pelos valores de uma métrica não necessariamente são os mais críveis quando outra métrica é utilizada.

Os resultados obtidos neste capítulo mostram que os comportamentos de concordância entre os trabalhadores são diversos. As métricas de credibilidade baseadas em concordância servem ao objetivo de dar proeminência a diferentes comportamentos que devem ser considerados pelos usuários. A interpretação que se faz dos valores de credibilidade e o uso deles em uma tomada de decisão dependem das características do projeto e dos interesses dos usuários. Os próximos capítulos exploram esses aspectos ao *(i)* analisar a relação entre credibilidade, engajamento e dificuldade das tarefas e ao *(ii)* aplicar as métricas de credibilidade na replicação de tarefas.

Capítulo 5

Relações entre o Engajamento e Credibilidade dos Trabalhadores e a Dificuldade das Tarefas

Além de delinear os comportamentos de engajamento e de credibilidade dos trabalhadores ao ofertarem poder cognitivo em projetos de computação por humanos, também é importante conhecer em que medida tais comportamentos se inter-relacionam e em que medida eles são afetados pelas características de dificuldade das tarefas. Nesse sentido, este capítulo descreve a pesquisa feita para investigar essas relações.

A pesquisa parte da investigação de duas questões principais, que são: *(i)* em que medida os valores de dificuldade das tarefas que os trabalhadores executam no projeto se relacionam com os valores de engajamento e de credibilidade que eles exibem e *(ii)* em que medida as características de engajamento manifestadas pelos trabalhadores se inter-relacionam com as características de credibilidade exibidas por eles. Esses estudos são conduzidos relacionando-se as métricas de engajamento, as métricas de credibilidade e a métrica de dificuldade propostas e discutidas nos capítulos anteriores.

5.1 Materiais e Métodos de Avaliação

O estudo das relações entre métricas de engajamento, credibilidade e dificuldade é realizado usando bases de dados obtidas de projetos reais de computação por humanos. Tais projetos

são discutidos na próxima seção. Em seguida, os métodos utilizados na análise das relações são detalhados.

5.1.1 Descrição dos Projetos Estudados

O estudo da relação entre engajamento, credibilidade e dificuldade requer que existam disponíveis na base de dados informações do instante de tempo em que os eventos ocorrem (para medição do engajamento) e das respostas providas pelos trabalhadores para as tarefas executadas (para medição da dificuldade e credibilidade). Das bases de dados apresentadas e estudadas nos capítulos anteriores, quatro possuem essas informações, são elas: Julgamento de Fatos, Análise de Sentimentos, Sun4All e Cell Spotting. Essas bases de dados são utilizadas no estudo apresentado neste capítulo. A base de dados do projeto Julgamento de Fatos, por não possuir informação temporal de execução de tarefas, não é utilizada na análise de relações que envolvam as medidas de engajamento.

5.1.2 Relação entre Engajamento e Dificuldade

Nos projetos estudados, as tarefas foram alocadas aos trabalhadores sem qualquer controle prévio da dificuldade. Em uma mesma sessão de trabalho, trabalhadores podem ter executado tarefas de diferentes graus de dificuldade. Em vista disso, não há como estabelecer uma relação sobre se os trabalhadores são mais engajados em tarefas fáceis ou em tarefas difíceis, pois eles foram expostos aos mais diversos tipos de tarefas durante o mesmo período de engajamento. Pode-se, entretanto, verificar a existência de uma relação média. Ou seja, analisar se a média do grau de dificuldade das tarefas às quais o trabalhador é exposto se relaciona com os valores de alguma de suas métricas de engajamento. Seja H_w um multiconjunto que contém os valores de dificuldades das tarefas executadas pelo trabalhador w. A dificuldade média experimenta pelo trabalhador w é calculada pela Equação 5.1.

$$\overline{h_w} = \frac{\sum\limits_{h \in H_w} h}{|H_w|} \tag{5.1}$$

A relação entre engajamento e dificuldade média é analisada em dois cenários: análise por trabalhador e análise no conjunto de todos os trabalhadores. Na análise por trabalha-

dor, ao fim de cada dia ativo, calculam-se as métricas de engajamento e a dificuldade média. Mede-se então, a correlação de Spearman entre cada métrica de engajamento e dificuldade média nos valores apresentados por cada trabalhador. Assim, para cada relação estudada, tem-se um valor de correlação por trabalhador. Analisa-se apenas trabalhadores que apresentam pelo menos 5 dias ativos. A métrica de engajamento duração relativa da atividade não se aplica nessa análise dado que a análise é feita considerando os dias anteriores à conclusão do projeto.

No segundo cenário, analisa-se a relação existente no conjunto de todos trabalhadores. Calculam-se as métricas de engajamento e a dificuldade média experimentada por cada trabalhador ao final do projeto. Mede-se então, a correlação de Spearman entre cada métrica de engajamento e a dificuldade média no conjunto de todos os trabalhadores. Assim, para cada relação estudada, tem-se um valor de correlação para todo o projeto. Nessa análise ao fim do projeto também se investiga como os diferentes perfis de engajamento diferem entre si em termos da distribuição da dificuldade média exibida pelos trabalhadores em cada perfil.

5.1.3 Relação entre Credibilidade e Dificuldade

A partir dos dados dos projetos estudados, é possível estimar o grau de dificuldade das tarefas e calcular a credibilidade dos trabalhadores para cada grau de dificuldade. Existem duas importantes questões a serem investigadas usando esses dados. A primeira questão é em que medida o desvio de credibilidade entre diferentes trabalhadores varia com a dificuldade das tarefas. Por exemplo, os trabalhadores são mais parecidos em termos de credibilidade em tarefas com baixo grau de dificuldade? Nessa análise, para cada grau de dificuldade, mede-se o desvio padrão da credibilidade dos trabalhadores usando as quatro métricas propostas.

A segunda questão é em que medida a credibilidade de um trabalhador varia com o grau de dificuldade da tarefa. Por exemplo, os trabalhadores são mais críveis em tarefas fáceis do que em tarefas difíceis? Dois cenários de análise são considerados: análise por trabalhador e análise no conjunto de todos os trabalhadores. Na análise por trabalhador, calculam-se as métricas de credibilidade para cada grau de dificuldade e mede-se a correlação de Spearman existente entre os valores de credibilidade e de dificuldade. Analisam-se apenas trabalhadores que executaram tarefas de pelo menos 5 graus de dificuldade diferentes. Na análise com todos os trabalhadores, por sua vez, ao final do projeto, calculam-se as métricas de credibi-

lidade para cada trabalhador em cada grau de dificuldade. Mede-se então, a correlação entre os valores de cada métrica de credibilidade e os graus de dificuldade no conjunto de todos os trabalhadores.

5.1.4 Relação entre Engajamento e Credibilidade

O estudo dessa relação visa elucidar em que medida uma métrica de engajamento se relaciona com uma métrica de credibilidade. Por exemplo, trabalhadores que apresentam alta taxa de atividade também apresentam alta concordância ponderada? Dado que a credibilidade é calculada por grau de dificuldade, a análise desse tipo de relação deve levar em conta a dificuldade.

Para analisar essa relação é utilizado o método de efeitos mistos (PINHEIRO; BATES, 2000). Por esse método, considera-se a existência de fatores fixos e fatores aleatórios. Os fatores fixos são as variáveis cuja relação se deseja estudar. Já os fatores aleatórios são variáveis cuja variação não é de interesse imediato, mas que podem afetar a relação entre os fatores fixos. No caso deste estudo, os fatores fixos são as métricas de engajamento e as métricas de credibilidade. O fator aleatório, por sua vez, é a dificuldade das tarefas.

A análise consiste em uma regressão multivariável na qual as métricas de engajamento são utilizadas para explicar cada métrica de credibilidade. A dificuldade é modelada como um fator aleatório que determina o intercepto da relação. Duas análises são conduzidas. Primeiro, investiga-se em que medida a dificuldade realmente é importante como fator aleatório nessa relação. Isso é analisado comparando um modelo que usa a dificuldade e um modelo que não a usa. A comparação dos modelos é feita usando análise de variância (ANOVA) que diz se há ganho de verossimilhança quando a dificuldade é utilizada. Segundo, analisa-se o coeficiente de regressão de cada métrica de engajamento explicando uma métrica de credibilidade. Esse coeficiente diz o poder explicativo de cada métrica de engajamento.

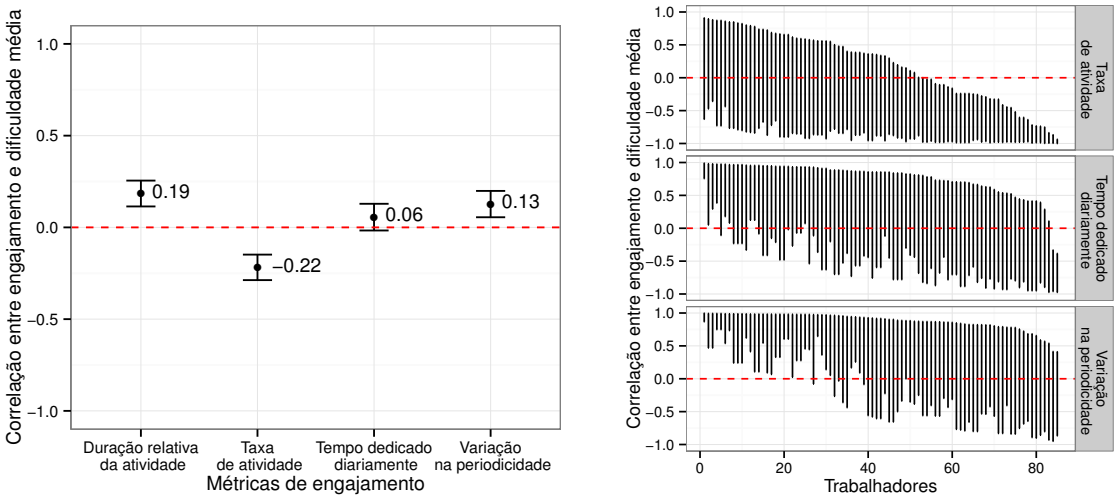
Todas as análise são realizas com o estado de engajamento e de credibilidade dos trabalhadores ao término do projeto. As métricas de engajamento são normalizadas utilizando o z-score.

5.2 Apresentação e Análise dos Resultados

Nesta seção, primeiro analisa-se a relação entre as métricas de engajamento dos trabalhadores e a dificuldade média das tarefas executadas por eles. Em seguida, analisa-se a relação entre as métricas de credibilidade e de dificuldade. Finalmente, discutem-se os resultados das relações entre métricas de engajamento e métricas de credibilidade.

5.2.1 Engajamento em Face da Dificuldade

Agora, verifica-se se a média ponderada do grau de dificuldade das tarefas às quais os trabalhadores foram expostos se relaciona com as características de engajamento que eles apresentam. Os resultados obtidos nos projetos Análise de Sentimentos, Cell Spotting e Sun4All são apresentados na Figuras 5.1, 5.2, e 5.3, respectivamente. Em cada figura, tem-se a correlação entre as métricas de engajamento e a dificuldade média no conjunto de trabalhadores ao fim do projeto e a correlação entre as métricas de engajamento e a dificuldade média percebida por trabalhador ao longo dos dias em que ele esteve ativo no projeto.

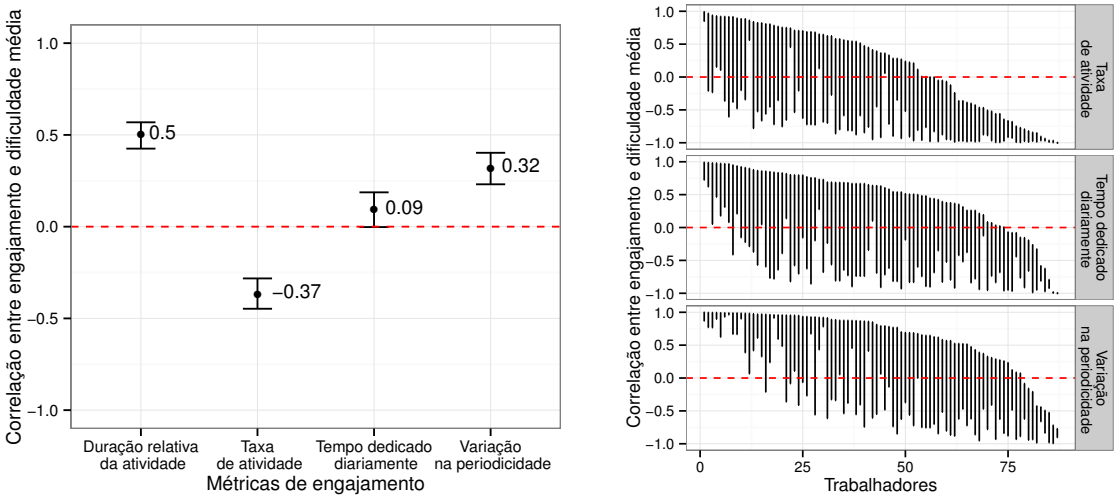


(a) Correlação no conjunto de trabalhadores ao fim do projeto

(b) Correlação por trabalhador ao longo dos dias ativos no projeto

Figura 5.1: Correlações entre métricas de engajamento e de dificuldade média percebida pelos trabalhadores no projeto Análise de Sentimentos. Mostram-se intervalos para um nível de confiança estatística de 95%. São significantes as correlações cujo intervalo não inclui o valor 0.

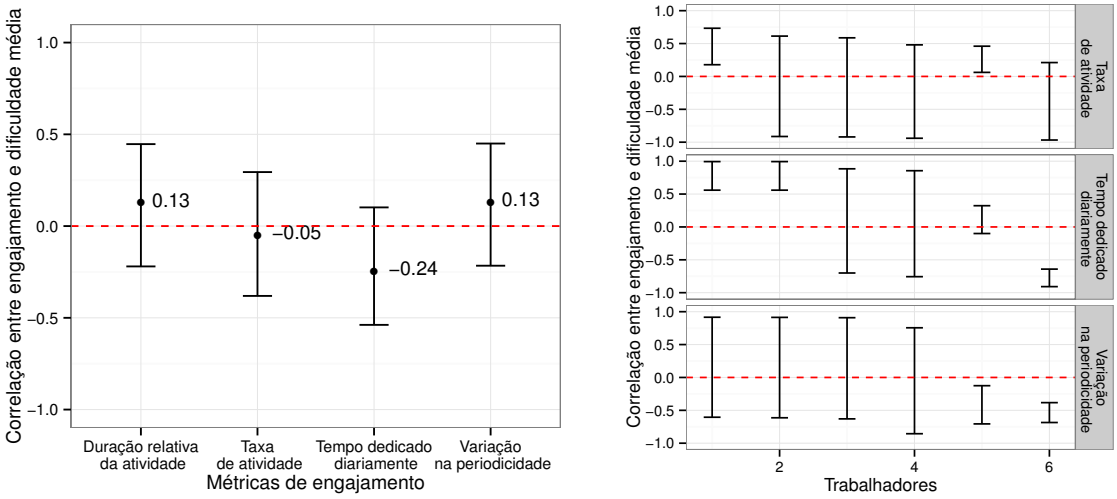
Quando considerado o conjunto de trabalhadores, observa-se a tendência de que trabalha-



(a) Correlação no conjunto de trabalhadores ao fim do projeto

(b) Correlação por trabalhador ao longo dos dias ativos no projeto

Figura 5.2: Correlações entre métricas de engajamento e de dificuldade média percebida pelos trabalhadores no projeto Cell Spotting. Mostram-se intervalos para um nível de confiança estatística de 95%. São significantes as correlações cujo intervalo não inclui o valor 0.



(a) Correlação no conjunto de trabalhadores ao fim do projeto

(b) Correlação por trabalhador ao longo dos dias ativos no projeto

Figura 5.3: Correlações entre métricas de engajamento e de dificuldade média percebida pelos trabalhadores no projeto Sun4All. Mostram-se intervalos para um nível de confiança estatística de 95%. São significantes as correlações cujo intervalo não inclui o valor 0.

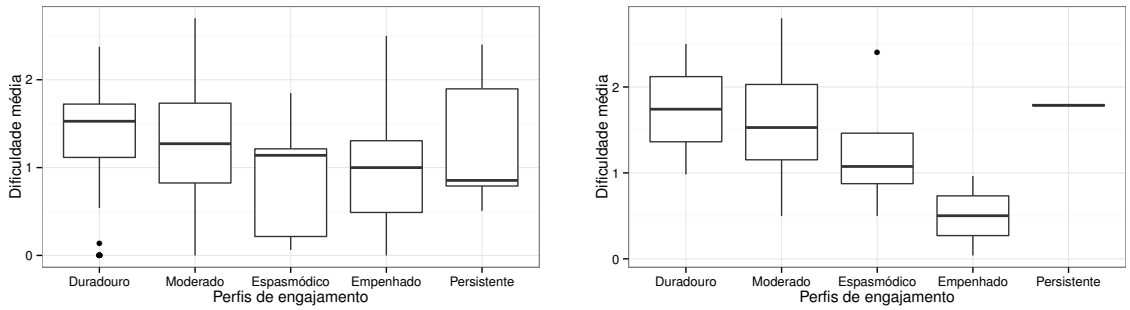
dores que foram expostos a tarefas com maior dificuldade média apresentam maior duração relativa da atividade (Cell Spotting, ρ=0,5; Análise de Sentimentos, ρ=0,19), maior variação na periodicidade (Cell Spotting, ρ=0,32; Análise de Sentimentos, ρ=0,13) e menor taxa de atividade (Cell Spotting, ρ=-0,37; Análise de Sentimentos, ρ=-0,22). Isso indica que *trabalhadores que executaram tarefas em média mais difíceis permaneceram no projeto durante um período de tempo maior, mas com menor atividade e maior variação na periodicidade dentro desse período*. No projeto Sun4All (Fig. 5.3(a)), entretanto, não se obteve resultados com significância estatística, provavelmente, em razão do baixo número de trabalhadores nesse projeto.

No que se refere à correlação entre as métricas de engajamento e a dificuldade média percebida por cada trabalhador ao longo dos dias ativos no projeto, observa-se que em todos os projetos há grande variação nas correlações apresentadas pelos trabalhadores (Análise de Sentimentos, Figura 5.1(b); Cell Spotting, Figura 5.2(b); e Sun4All, Figura 5.3(b)). Em todas as métricas, alguns trabalhadores apresentam correlação positiva, outros trabalhadores apresentam correlação negativa e muitos não apresentam correlação significativa. Isso indica que a relação entre as métricas de engajamento e dificuldade média é muito dependente das características individuais dos trabalhadores. Não há um comportamento geral que possa ser apontado. Isso também indica que os trabalhadores reagem de forma bem diferente quando expostos à dificuldade e, nesse caso, as diferenças individuais precisam ser consideradas.

Finalmente, analisou-se a dificuldade média experimenta por cada trabalhador nos diferentes perfis de engajamento. Os resultados são apresentados na Figura 5.4 para os projetos Cell Spotting e Sun4All. Nessa figura, cada caixa indica a distribuição das dificuldades médias experimentadas pelos trabalhadores no perfil indicado no eixo horizontal. Considerando a mediana, observa-se uma ordem de maior para menor dificuldade experimentada por trabalhadores que exibem os perfis de engajamento duradouro, moderado, espasmódico e empenhado, respectivamente.

5.2.2 Credibilidade em Face da Dificuldade

Quando se considera o cálculo de credibilidade por grau de dificuldade de tarefa, tem-se diversos valores de credibilidade para cada trabalhador, sendo um para cada grau de dificuldade. Uma importante questão que surge nesse contexto é, em que medida as credibilidades



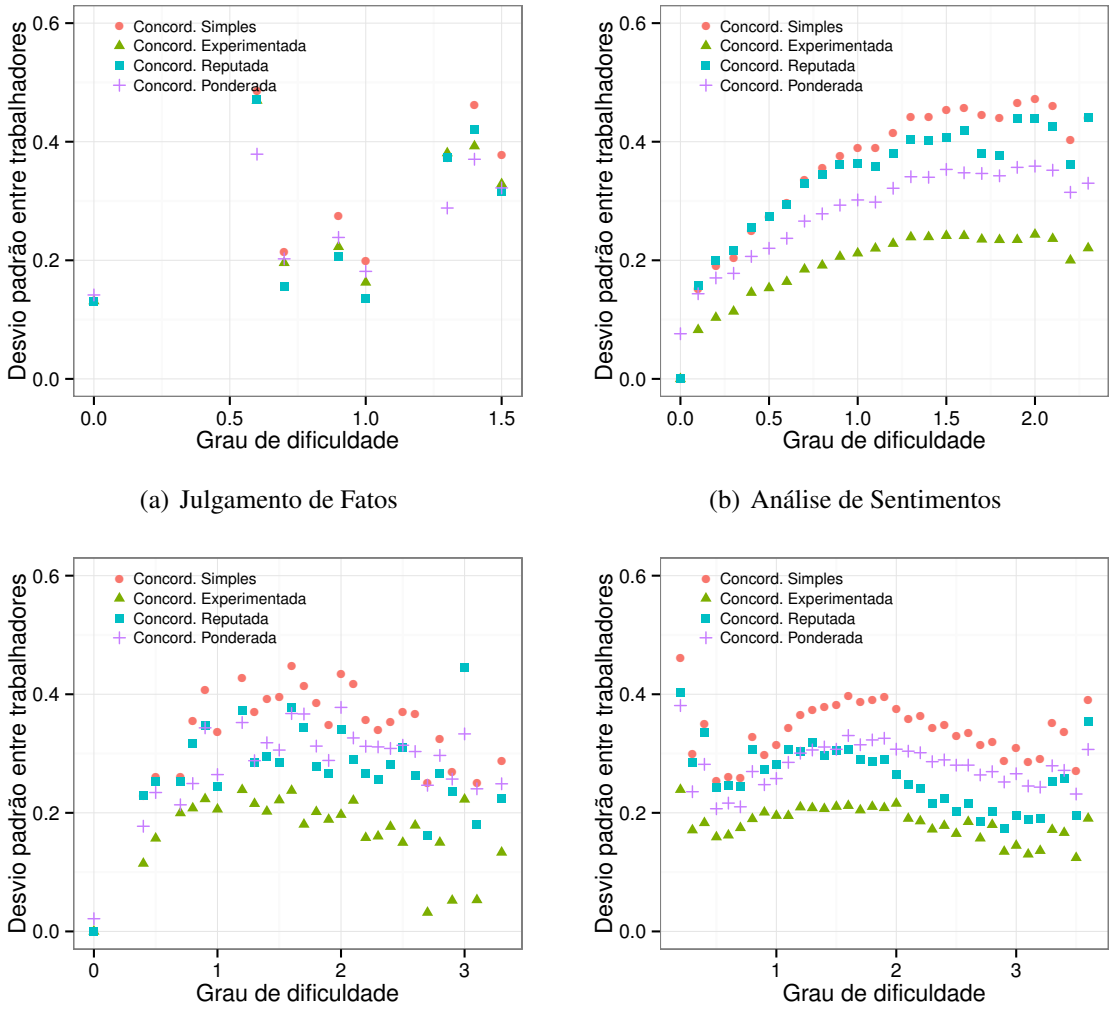
(a) Cell Spotting (b) Sun4All

Figura 5.4: Dificuldade média percebida por trabalhadores que exibem os perfis de engajamento Empenhado, Espasmódico, Persistente, Duradouro e Moderado. Mostram-se resultados obtidos nos projetos (a) Cell Spotting e (b) Sun4All.

dos trabalhadores variam no mesmo grau de dificuldade. Por exemplo, todos os trabalhadores são igualmente críveis em tarefas com grau de dificuldade 0? Nesse caso, pode-se calcular a credibilidade de cada trabalhador em cada grau de dificuldade e medir o desvio padrão desses valores. Esse desvio diz em que medida os trabalhadores são diferentes. Se o desvio padrão é 0, todos os trabalhadores exibem igual credibilidade no dado grau de dificuldade. De outro modo, quanto maior é o desvio padrão maior é a variação da credibilidade dos trabalhadores no dado grau de dificuldade.

Os resultados dessa análise são apresentados na Figura 5.5 para os projetos Julgamento de Fatos (Fig. 5.5(a)), Análise de Sentimentos (Fig. 5.5(b)), Sun4All (Fig. 5.5(c)) e Cell Spotting (Fig. 5.5(d)). Esses resultados mostram que *os trabalhadores tendem a apresentar credibilidade mais diferente entre si em tarefas com grau de dificuldade moderada ou alta*. No projeto Julgamento de Fatos existem poucos graus de dificuldade de tarefas e não é clara uma tendência. No projeto Análise de Sentimentos, o desvio de credibilidade entre os trabalhadores aumenta quando a dificuldade da tarefa aumenta e depois se estabiliza. Já nos projetos Sun4All e Cell Spotting, observa-se um comportamento não monotônico. Há maior desvio em graus de dificuldade moderados. Quando se observa as métricas de credibilidade, observa-se que os maiores desvios ocorrem com a métrica de concordância simples e os menores desvios ocorrem com a métrica de concordância experimentada.

Para cada trabalhador, também se pode verificar em que medida a credibilidade dele varia em diferentes graus de dificuldade. Nesse caso, pode-se, para cada trabalhador, calcular a credibilidade dele em tarefas com cada grau de dificuldade e o desvio padrão desses valores



(a) Julgamento de Fatos

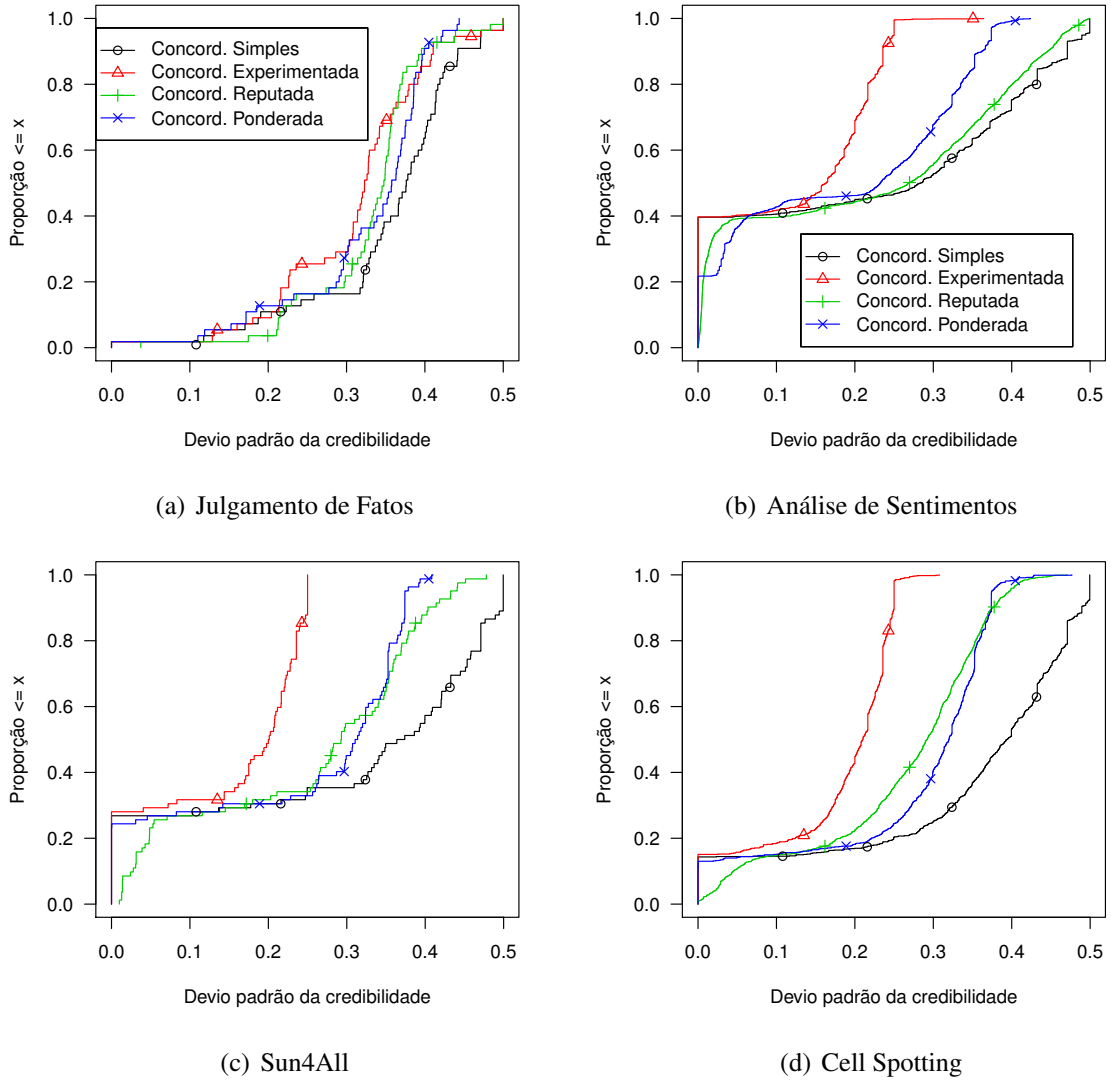
(b) Análise de Sentimentos

(c) Sun4All

(d) Cell Spotting

Figura 5.5: Distribuição dos desvios padrões na credibilidade do conjunto de trabalhadores que executaram tarefas em cada grau de dificuldade. Mostram-se resultados obtidos nos projetos (a) Julgamento de Fatos, (b) Análise de Sentimentos, (c) Sun4All e (d) Cell Spotting.

de credibilidade diz em que medida a credibilidade do trabalhador varia ao longo de diferentes graus de dificuldade. Se o desvio padrão é 0, a credibilidade não varia com a dificuldade e, de outro modo, quando maior é o desvio padrão maior é a variação da credibilidade. A Figura 5.6 mostra as distribuições dos desvios padrões nas credibilidade dos trabalhadores nos projetos Julgamento de Fatos (Fig. 5.6(a), Análise de Sentimentos (Fig. 5.6(b)), Sun4All (Fig. 5.6(c)) e Cell Spotting (Fig. 5.6(d)).



(a) Julgamento de Fatos

(b) Análise de Sentimentos

(c) Sun4All

(d) Cell Spotting

Figura 5.6: Distribuição dos desvios padrões na credibilidade de cada trabalhador ao longo de diferentes graus de dificuldade de tarefa. Mostram-se resultados obtidos nos projetos (a) Julgamento de Fatos, (b) Análise de Sentimentos, (c) Sun4All e (d) Cell Spotting.

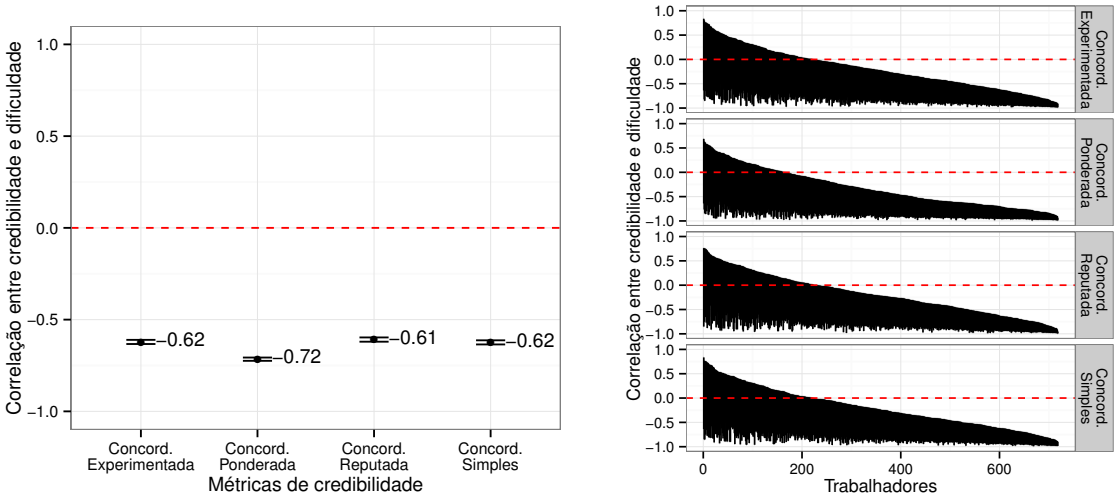
Existem dois pontos principais a serem destacados nesses resultados. O primeiro ponto é que a *maioria dos trabalhadores apresenta algum desvio na credibilidade ao longo de diferentes graus de dificuldade*. No projeto Análise de Sentimentos, em que há mais trabalhadores com desvio abaixo, existe 40% dos trabalhadores com desvio zero na métrica de concordância experimentada. Enquanto no projeto Julgamento de Fatos, no qual há menos trabalhadores com baixo desvio, 2% dos trabalhadores exibem desvio 0 na métrica de concordância ponderada. O segundo ponto a ser destacado é que, quando se analisa as distribuições como um todo, os menores desvios são observados na métrica de concordância experimentada, onde, portanto, tende a ocorrer melhor variação da credibilidade com a dificuldade. Por outro lado, os maiores desvios são observados na métrica de concordância simples, portanto é a métrica mais sensível ao grau de dificuldade da tarefa.

Pode-se agora analisar a correlação existente entre credibilidade dos trabalhadores e dificuldade percebida por eles. Os resultados se encontram apresentados nas Figuras 5.7, 5.8, e 5.9 para os projetos Análise de Sentimentos, Cell Spotting e Sun4All , respectivamente. Em cada figura, tem-se a correlação entre as métricas de credibilidade e os graus de dificuldade no conjunto de todos os trabalhadores e essa correlação por trabalhador. Em todos os projetos estudados, observa-se uma correlação negativa entre os valores de credibilidade e os graus de dificuldade. Isso indica uma associação de que *quanto maior a dificuldade das tarefas menos críveis os trabalhadores tendem a ser*. Essa associação também é observada quando se analisa as correlações por trabalhadores. A maioria dos trabalhadores apresenta uma correlação negativa significativa.

5.2.3 Inter-relações entre Engajamento e Credibilidade

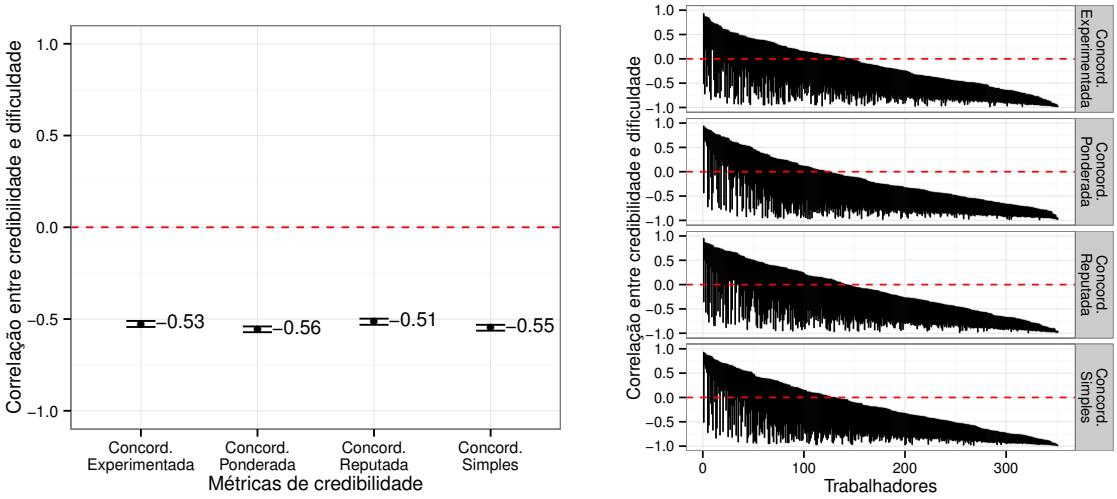
Além de compreender como os trabalhadores se engajam em computação por humanos e as suas credibilidades em tarefas com diferentes dificuldades, também importa conhecer eventuais relações que possam existir entre engajamento e credibilidade. Neste contexto, há dois pontos de interesse: (i) entender a importância da dificuldade das tarefas na análise da relação entre métricas de credibilidade e métricas de engajamento, e (ii) entender a força da relação entre essas métricas.

A Tabela 5.1 apresenta o ganho de verossimilhança de um modelo de regressão em que o intercepto varia com o grau de dificuldade em relação ao modelo de regressão em que o in-



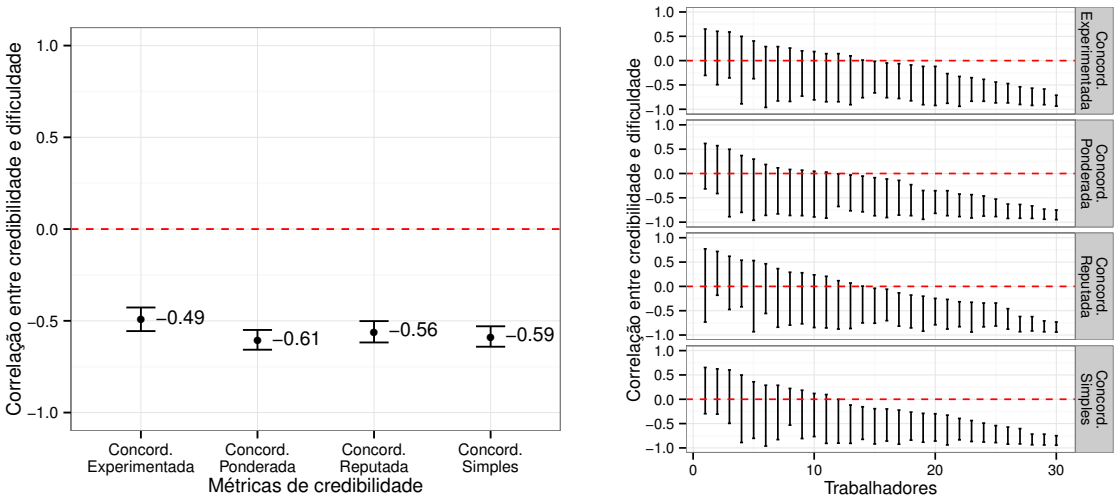
(a) Correlação no conjunto de todos os trabalhadores (b) Correlação por trabalhador

Figura 5.7: Correlações entre métricas de credibilidade e de dificuldade no projeto Análise de Sentimentos. Mostram-se intervalos para um nível de confiança estatística de 95%. São significantes as correlações cujo intervalo não inclui o valor 0.



(a) Correlação no conjunto de todos os trabalhadores (b) Correlação por trabalhador

Figura 5.8: Correlações entre métricas de credibilidade e de dificuldade no projeto Cell Spotting. Mostram-se intervalos para um nível de confiança estatística de 95%. São significantes as correlações cujo intervalo não inclui o valor 0.



(a) Correlação no conjunto de todos os trabalhadores

(b) Correlação por trabalhador

Figura 5.9: Correlações entre métricas de credibilidade e de dificuldade no projeto Sun4All. Mostram-se intervalos para um nível de confiança estatística de 95%. São significantes as correlações cujo intervalo não inclui o valor 0.

Tabela 5.1: Ganho de verossimilhança do modelo de regressão com intercepto variável com o grau de dificuldade das tarefas em relação ao modelo de regressão com intercepto fixo.

	Concordância simples	**Concordância Experimentada**	**Concordância Reputada**	**Concordância Ponderada**
Análise de Sentimentos	1654,47	1545,68	1576,57	2050,63
Sun4All	200,56	174,46	209,73	197,50
Cell Spotting	4133,46	4244,23	3735,41	4259,61

Nota: Em todos os casos o ganho é significante com p-valor < 0.0001.

tercepto é fixo. Em todos os projetos, há ganho de verossimilhança ao se modelar a variação com o grau de dificuldade. Ou seja, todos os modelos de regressão são significativamente melhores. Isso indica que *a capacidade do engajamento dos trabalhadores de explicarem as credibilidades deles depende da dificuldade da tarefa*.

O segundo ponto a ser tratado é qual a capacidade das métricas de engajamento de explicarem as credibilidades dos trabalhadores. A Figura 5.10 mostra os coeficientes de regressão das métricas de engajamento explicando cada métrica de credibilidade. A significância, o valor, e a tendência do coeficiente são dependentes do projeto. No projeto Sun4All (Fig. 5.10(a)), observa-se uma tendência de que alta duração relativa da atividade e alta taxa de atividade estejam relacionados à menor credibilidade. No projeto Análise de Sentimentos (Fig. 5.10(b)), por sua vez, a taxa de atividade está positivamente relacionada com a credi-

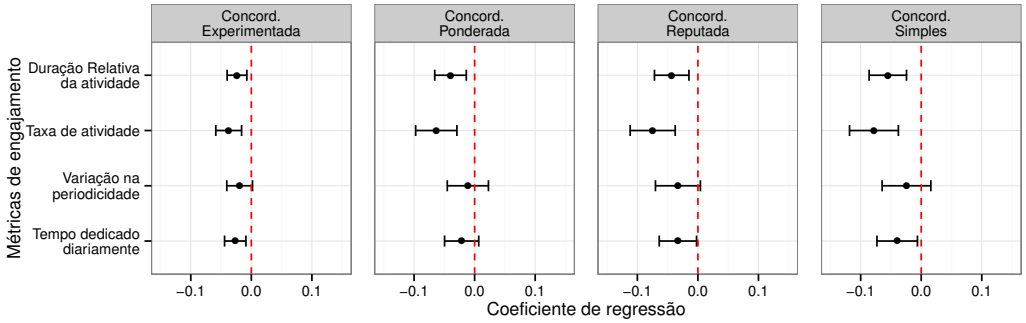
bilidade. Finalmente, no projeto Cell Spotting (Fig. 5.10(c)), a taxa de atividade e o tempo dedicado diariamente estão negativamente relacionados com a credibilidade, enquanto a variação na periodicidade está negativamente relacionada com a credibilidade. Pelos coeficientes, observa-se que a variação de 1 desvio padrão em uma métrica de engajamento explica uma variação de no máximo 0,1 no valor de credibilidade. Assim, embora significantes, os coeficientes são baixos. Ou seja, *os valores de engajamento explicam pouco os valores de credibilidade*.

5.3 Considerações Finais

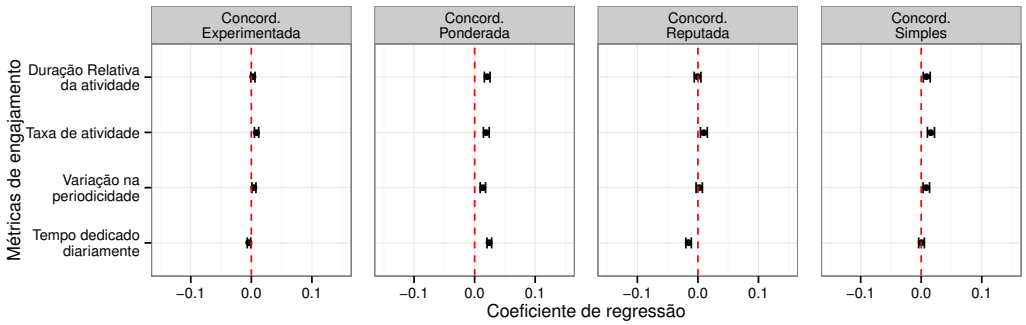
Neste capítulo, investigou-se as relações entre métricas de engajamento, métricas de credibilidade e a métrica de dificuldade das tarefas. Os principais resultados da análise dessas relações são:

- Em dois dos três projetos estudados, trabalhadores que executam tarefas em média mais difíceis apresentam maior período de engajamento, mas com menor atividade e maior variação na periodicidade dentro desse período;
- Trabalhadores tendem a apresentar credibilidades mais diferentes entre si em tarefas com grau de dificuldade moderada ou alta;
- A maioria dos trabalhadores apresenta algum desvio na credibilidade ao longo de diferentes graus de dificuldade. Quanto maior a dificuldade das tarefas, menos críveis eles tendem a ser;
- A relação entre o engajamento dos trabalhadores e a credibilidades deles depende da dificuldade da tarefa;
- Os valores de engajamento dos trabalhadores explicam pouco os valores de credibilidade deles.

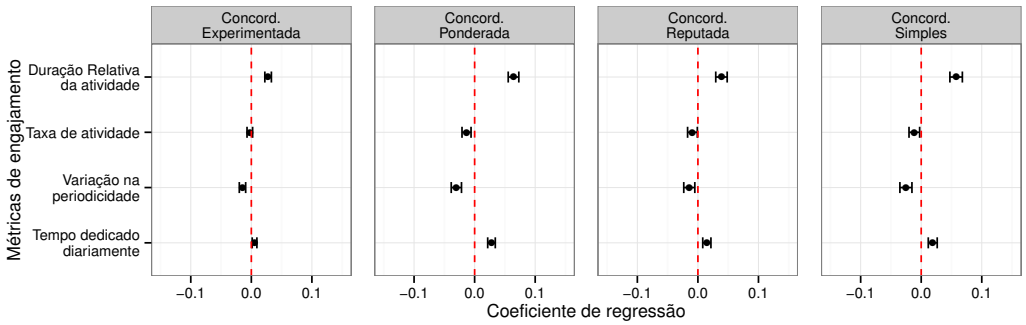
Uma importante lição a ser pontuada na pesquisa reportada neste capítulo é que não é clara a relação entre engajamento e dificuldade e a relação entre engajamento e credibilidade. Essas relações tendem a ser dependentes de peculiaridades dos trabalhadores e dos



(a) Sun4All



(b) Análise de Sentimentos



(c) Cell Spotting

Figura 5.10: Coeficientes de regressão entre métricas de engajamento e de credibilidade. Mostram-se resultados obtidos nos projetos (a) Sun4All, (b) Análise de Sentimento e (c) Cell Spotting. São apresentados intervalos para um nível de confiança estatística de 95%. São significantes os coeficientes cujo intervalo não inclui o valor 0.

projetos. Não há um padrão geral que possa ser apontado. No entanto, a relação entre dificuldade e credibilidade é bastante clara. Os valores de credibilidade dos trabalhadores são fortemente correlacionados com os valores de dificuldade das tarefas. A diferença de credibilidade entre trabalhadores também é muito influenciada pela dificuldade das tarefas. Dessa forma, em computação por humanos, o estudo da credibilidade dos trabalhadores não deve ser desassociado do estudo da dificuldade das tarefas. Esse resultado é considerado na pesquisa descrita no próximo capítulo na qual se propõe e se avalia um algoritmo de replicação de tarefas que faz uso de informações de credibilidade dos trabalhadores e de dificuldade de tarefas.

Capítulo 6

Replicação de Tarefas em Computação por Humanos

Conhecer a oferta de poder cognitivo em projetos de computação por humanos por meio da caracterização do engajamento e da credibilidade dos trabalhadores, tal como realizado nos capítulos anteriores, é um grande avanço na análise do desempenho de tais projetos. No entanto, um passo seguinte a essa compreensão da oferta de poder cognitivo é desenvolver projetos de computação por humanos capazes de usar esse poder cognitivo de forma otimizada. O conhecimento do engajamento e da credibilidade dos trabalhadores pode informar o desenvolvimento de diversas estratégias que visem melhorar o desempenho de projetos de computação por humanos. Neste capítulo, descreve-se a pesquisa feita para otimizar o desempenho de projetos de computação por humanos por meio da replicação de tarefas que leva em conta informações da credibilidade dos trabalhadores e da dificuldade das tarefas.

Atualmente, na ausência de conhecimento sobre a credibilidade dos trabalhadores, a abordagem mais adotada para tratar incertezas nas respostas providas por eles é a obtenção e agregação de respostas redundantes (HOVY et al., 2013; SHESHADRI; LEASE, 2013; PONCIANO et al., 2014a). Nessa abordagem, cada tarefa é executada por diversos trabalhadores. Após as respostas redundantes serem obtidas dos trabalhadores, aplica-se um algoritmo de agregação com tratamento de incertezas e se obtém a resposta final para a tarefa. O algoritmo mais simples e mais utilizado na prática é o voto majoritário, em que se considera correta a resposta mais frequente no conjunto de respostas redundantes (SHESHADRI; LEASE, 2013). A abordagem de redundância seguida de agregação gera diversos problemas. Um dos principais

problemas é a definição da quantidade de respostas redundantes que terão que ser obtidas para cada tarefa. Se ela for subestimada, compromete-se a acurácia da resposta obtida na agregação. Por outro lado, se ela for superestimada, há um excesso de trabalho redundante e consequente desperdício de poder cognitivo provido pelos trabalhadores.

A pesquisa aqui reportada parte de uma discussão conceitual sobre o que é replicação de tarefas em sistemas computacionais e de como tal replicação pode ser feita no contexto de projetos de computação por humanos de modo a melhorar o uso do poder cognitivo disponível. A partir desse entendimento, propõe-se uma estratégia de replicação de tarefas cujo propósito principal é definir de forma adaptativa e otimizada o nível de redundância que precisa ser utilizado em cada tarefa enquanto permite parametrizar requisitos de interesse dos usuários como urgência, métrica de credibilidade a ser utilizada e nível de credibilidade requerida nas respostas.

Nas seções seguintes, primeiro apresenta-se uma contextualização de replicação de tarefas (Seção 6.1). Após isso, propõe-se um algoritmo de replicação de tarefas para projetos de computação por humanos (Seção 6.2). Finalmente, apresenta-se a avaliação realizada usando dados de projetos reais, detalhando-se os materiais e métodos de avaliação (Seção 6.3) e os resultados obtidos (Seção 6.4).

6.1 Fundamentos de Replicação de Tarefas

Busca-se nesta seção prover uma base conceitual sobre replicação de tarefas por meio da análise do conceito de replicação, os propósitos para os quais replicação geralmente é utilizada, os tipos de replicação em sistemas computacionais e as definições de grau de replicação e agregação de respostas de diferentes réplicas.

6.1.1 O que é Replicação?

Replicação é um mecanismo utilizado quando é importante se gerar algum tipo de redundância. Esse mecanismo tem inspiração na natureza. Por exemplo, replicação pode ser observada no processo de redundância genética, fenômeno comum nos seres vivos em que determinada função bioquímica é codificada em dois ou mais genes (KAFRI; LEVY; PILPEL, 2006). Tal redundância permite que as mutações ou defeitos em um desses genes tenham um

efeito menor sobre a aptidão do ser vivo, pois os outros genes podem assumir a função (KAFRI; LEVY; PILPEL, 2006).

Replicação é amplamente utilizada no projeto de sistemas computacionais. Em sistemas distribuídos (COULOURIS; DOLLIMORE; KINDBERG, 2005; TANENBAUM; STEEN, 2006), estratégias de replicação são amplamente utilizadas como forma de se prover redundância de dados, em que um mesmo dado é disponibilizado em diversos recursos, e redundância de computações, em que uma mesma computação é realizada diversas vezes. Em sistemas baseados na escolha social ou na concordância entre seres humanos (COHEN, 1960; KRIPPENDORFF, 1970; FLEISS; LEVIN; PAIK, 1981), replicação é utilizada como forma de se obter redundância de respostas ou julgamentos para elicitar preferências ou escolhas coletivas.

6.1.2 Propósito da Replicação

Replicação geralmente é realizada para se obter alta disponibilidade/desempenho do sistema ou para se tolerar falhas em algum dos seus componentes (COULOURIS; DOLLIMORE; KINDBERG, 2005, p. 603).

Replicação como forma de aumentar o desempenho ocorre, por exemplo, em implementações de *caching* de dados. Pelo mecanismo de *caching*, dados são replicados de um recurso no qual o acesso aos dados é mais lento para outro recurso no qual o dado possa ser acessado com maior velocidade (COULOURIS; DOLLIMORE; KINDBERG, 2005; PONCIANO; ANDRADE; BRASILEIRO, 2013). Replicações de computações também podem ser utilizadas para aumentar o tempo de resposta de um sistema. Por exemplo, em ambientes compostos por recursos que podem apresentar diferentes poderes computacionais, uma computação pode ser replicada para diversos recursos com o propósito de que algum desses recursos conclua a computação de forma mais rápida e, assim, reduza o tempo de resposta do sistema (SILVA; CIRNE; BRASILEIRO, 2003; CIRNE et al., 2007).

Replicação como forma de tolerar falhas, por sua vez, geralmente distingue três tipos principais de redundância: redundância de hardware, redundância de software e redundância no tempo (JALOTE, 1994). Redundância de hardware consiste em realizar uma computação (ou manter o estado dela) em diferentes recursos computacionais. Isso permite tolerar falhas em recursos. Redundância de software significa, por exemplo, estarem disponíveis diferentes versões de compiladores, sistemas operacionais e outros softwares ou informações nos

quais as computações se baseiam. Isso permite tolerar falhas em, por exemplo, versões de softwares. Finalmente, redundância no tempo consiste em repetir (ou reiniciar) uma mesma computação diversas vezes ao longo do tempo. Isso permite tolerar falhas esporádicas em razão, por exemplo, do estado de contenção do recurso em um dado momento.

É importante ressaltar que a distinção entre os hardwares que realizam a computação e os softwares nos quais ela se baseia não é tão clara em computação por humanos como é em computação por máquinas. Embora o cérebro humano possa ser visto como um hardware e a mente humana possa ser vista como um conjunto de softwares (SCHWARTZ; BEGLEY, 2009), em computação por humanos ambos não podem ser tão facilmente separados de modo a distinguir os efeitos de cada um deles como pode ser feito em computação por máquinas. Neste trabalho, trata-se o sistema cognitivo humano sem distinção entre software e hardware.

6.1.3 Tipos de Replicação

Muitas das primeiras abordagens que empregaram replicação de tarefas em sistemas distribuídos são baseadas nos conceitos de replicação ativa e replicação passiva (SCHNEIDER, 1990). Replicação é dita ativa quando cada réplica de uma tarefa é executada integralmente por recursos diferentes, cada uma partindo de um mesmo estado inicial até o mesmo estado final da tarefa. De outro modo, a replicação é dita passiva quando recursos adicionais são utilizados para manter uma cópia (*backup*) do estado da tarefa na medida em que ela é executada por um recurso principal. Essa cópia mantém as computações já realizadas e ela pode ser utilizada para que a tarefa não precise ser re-executada integralmente caso uma falha ocorra durante sua execução.

O conceito de replicação passiva de tarefas de computação por humanos é possível em tarefas longas, nas quais o trabalhador gera uma sequência de resultados para uma mesma tarefa. No entanto, tipicamente, tarefas de computação por humanos são de curta duração. A tarefa é executada de uma só vez e o trabalhador provê apenas uma resposta final. Não há estados intermediários para que cópias parciais sejam mantidas. Dessa forma, o conceito de replicação ativa se aplica melhor. Esse tipo de replicação tem sido utilizado em computação por humanos como forma de obter uma redundância de respostas. Tal redundância pode ser utilizada tanto para analisar opiniões/preferências de diversos trabalhadores quanto para tolerar falhas nas execuções.

Replicação também pode ser diferenciada em abordagens definidas como replicação no tempo e replicação no espaço (JALOTE, 1994). A replicação no tempo consiste em uma mesma computação ser replicada diversas vezes ao longo do tempo em um mesmo recurso. Por exemplo, uma mesma tarefa pode ser executada por um mesmo trabalhador diversas vezes ao longo do tempo. Isso permitira, por exemplo, tratar erros humanos não sistemáticos como esquecimento e deslizes, mas não teria eficácia no caso de ignorância e vieses cognitivos. Já a replicação no espaço se baseia no uso de diversos recursos nos quais itens são replicados. Esse tipo de replicação requer a existência de uma diversidade de recursos, mas também pode ser eficaz na tolerância a falhas.

6.1.4 Grau de Replicação e Agregação de Respostas

Quando se utiliza replicação de tarefas em diferentes recursos, o *grau de replicação* é a quantidade de recursos diferentes nos quais a tarefa será replicada e o mecanismo de *agregação de respostas* é a forma como se elegerá a resposta final para a tarefa a partir da respostas providas pelos diferentes recursos nos quais as réplicas foram executadas.

Com o uso de replicação, as respostas obtidas no sistema podem ser representadas em uma matriz S de dimensão $|W| \times |T|$, como na Figura 6.1 adaptada de Law e Ahn (2011). Considere W o conjunto de trabalhadores que executam tarefas em um projeto e T o conjunto composto pelas tarefas no projeto. Um trabalhador $w \in W$ ao executar uma tarefa $t \in T$ gera um resposta $s_{w,t}$. Assim, cada linha da matriz S corresponde a um trabalhador, cada coluna corresponde a uma tarefa da aplicação e cada valor é uma resposta gerada pelo correspondente trabalhador para a correspondente tarefa. Naturalmente, nessa matriz podem existir valores não definidos, pois nem todos os trabalhadores precisam executar todas as tarefas.

O problema tratado pela agregação de respostas é identificar a resposta correta em cada coluna da matriz S. A quantidade de respostas existentes em cada coluna é o grau de replicação utilizado na execução das tarefas, ou seja, a quantidade de trabalhadores diferentes que apresentaram respostas para a tarefa. Definir o nível de replicação adequado é um desafio inerente ao estudo de replicação. Aumentar o número de réplicas pode estar associado a um maior custo em termos, por exemplo, da quantidade de trabalhadores utilizados. Enquanto, por outro lado, reduzir o número de réplicas pode comprometer a acurácia da resposta obtida

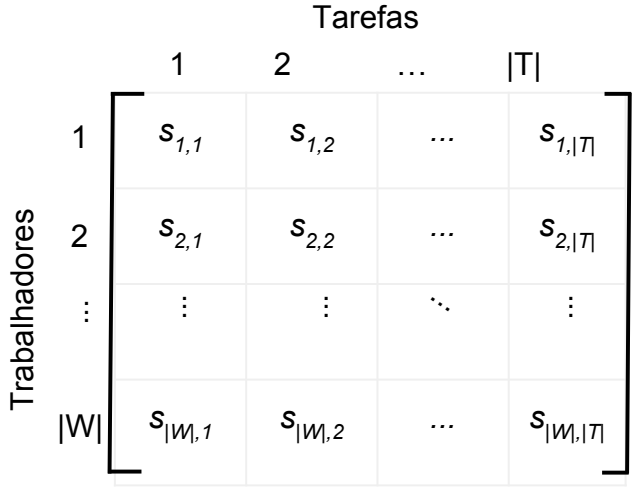


Figura 6.1: Estrutura de uma matriz composta por respostas geradas por diversos trabalhadores para diversas réplicas das tarefas de um projeto.

na agregação.

6.2 Algoritmo de Replicação Adaptativa baseada em Credibilidade

As métricas de engajamento e de credibilidade dos trabalhadores apresentadas e analisadas nos últimos capítulos permitem avaliar características de oferta de poder cognitivo pelos trabalhadores. Nesta seção, propõe-se um algoritmo de replicação cujo objetivo é definir o grau de replicação a ser utilizado em cada tarefa do projeto de modo a evitar excesso de redundância e conseguente desperdício de poder cognitivo.

O Algoritmo 1 apresenta a sequência de passos[1] computados na replicação de uma tarefa. Dada uma tarefa de computação por humanos t, o objetivo do algoritmo é obter uma resposta final para a tarefa e a credibilidade associada a essa resposta. Isso é realizado tentando gerar o mínimo possível de réplicas e considerando as seguintes restrições:

- Limite mínimo de credibilidade requerida na resposta final para a tarefa (variável $credRequ$). É um valor decimal entre 0 e 1 que indica o nível de credibilidade desejada para que a resposta final obtida pelo algoritmo seja considerada crível.

[1]Para manter a simplicidade e a clareza dos passos de computação, o algoritmo é apresentado sem otimizações de desempenho nas computações realizadas por computadores digitais.

- Limite máximo de réplicas (variável $maxRepl$). Trata-se de um número inteiro, positivo e maior que 0 que indica o limite máximo de réplicas que podem ser geradas pelo algoritmo de replicação.

- Urgência da execução (variável $urge$). É um valor decimal entre 0 e 1 que indica o nível de urgência em se obter uma resposta final para a tarefa. Quando maior a urgência, maior o paralelismo permitido na execução das réplicas e menor o espaço para otimizar o grau de replicação.

No Algoritmo 1, calcula-se o grau de paralelismo (variável $numReplPorTurno$, linha 3) em função do limite máximo de réplicas que pode ser gerado e da urgência. Quando o parâmetro de urgência assume o valor 1, todas as réplicas devem ser geradas ao mesmo tempo. Nesse caso, o algoritmo não tem possibilidade de otimizar o número de réplicas que serão geradas. Por outro lado, quando a urgência é 0, a geração das réplicas é sequencial, ou seja, gera-se uma réplica de cada vez.

Algoritmo 1: Replicação Adaptativa Baseada em Credibilidade.

Entrada: Tarefa t, Credibilidade requerida $credRequ$, Máximo de réplicas $maxRepl$, Urgência $urge$
Saída: Resposta final para a tarefa $respGrupo$, Credibilidade da resposta final $credGrupo$;

```
1  contRepl ← 0;                                          // Contador de réplicas
2  S_t ← {}; /* Mapa cujas chaves são respostas e os valores são listas que
   mantêm os identificadores dos trabalhadores que proveram as respostas
   indicadas nas chaves.  */
3  numReplPorTurno ← max(maxRepl * urge, 1);
4  repita
5      numReplNesteTurno ← min(numReplPorTurno, maxRepl − contRepl);
6      atribuiReplicas(numReplNesteTurno, t, S_t);        /* gera numReplNesteTurno
       réplicas da tarefa t, obtém as respostas dos trabalhadores e as
       mantém no mapa S_t */
7      G ← calculaCredibilidadeDosTrabalhadores(S_t); /* Definido no Algoritmo 2 */
8      respGrupo, credGrupo ← pegaGrupoTrabalhadoresMaisCrivel(G);  /* Definido no
       Algoritmo 3 */
9      contRepl ← contRepl + numReplNesteTurno;
10 até credGrupo ≥ credRequ or contRepl = maxRepl ;
11 retorna respGrupo, credGrupo;
```

A atribuição das réplicas geradas aos trabalhadores (procedimento $atribuiReplica()$, linha 6 do Algoritmo 1) é uma chamada bloqueante à Interface de Programação de Aplicações (API, do inglês *Application Programming Interface*) do sistema de computação por humanos. O comportamento desse método varia com o sistema. Ele pode consistir em, por exemplo: (a) disponibilizar as réplicas da tarefa no quadro de trabalho para que elas possam ser

escolhidas pelos trabalhadores, ou (b) escalonar as réplicas da tarefa para os trabalhadores disponíveis. Uma vez que as réplicas são executadas, o sistema atualiza o mapa S_t adicionando as respostas para as réplicas e os identificadores dos trabalhadores que as proveram. Nesse mapa, as chaves são respostas e os valores são listas que mantêm os identificadores dos trabalhadores que proveram as respostas indicadas nas chaves.

Com os identificadores dos trabalhadores, pode-se calcular suas credibilidades usando uma das métricas propostas no Capítulo 4. A credibilidade é calculada usando todo o histórico de respostas providas pelo trabalhador para as tarefas anteriores à tarefa corrente. Trabalhadores que ainda não executaram tarefas são iniciados com credibilidade igual a 0,5. No Algoritmo 1, a credibilidade de cada trabalhador é calculada pela função $calculaCredibilidadeDosTrabalhadores()$, chamada na linha 7. A implementação dessa função é apresentada no Algoritmo 2.

Algoritmo 2: Calcula as Credibilidade dos Trabalhadores.

Entrada: Mapa S_t;
Saída: Mapa G;

```
1 G ← {};      /* G é um mapa cujas chaves são respostas e os valores são
  listas que mantêm as credibilidades dos trabalhadores que proveram as
  respostas indicadas nas chaves.  */
2 para cada chave resposta em S_t faça
3     para cada trabalhador em S_t[resposta] faça
4         credDoTrab ← calculaCredibilidadeDoTrabalhador(S_t, trabalhador, t);
5         G[resposta].adiciona(credDoTrab);
6     fim
7 fim
8 retorna G;
```

Respostas iguais obtidas de diferentes trabalhadores podem ser agrupadas (Algoritmo 3). O número de grupos, que varia com a diversidade de respostas geradas pelos trabalhadores, é definido como n. Seja o grupo G_a o multiconjunto das credibilidades dos trabalhadores que proveram a resposta a. Pode-se então calcular a credibilidade de cada grupo de respostas. Essa credibilidade é definida como a probabilidade de resposta a provida por um determinado grupo de trabalhadores de credibilidades G_a ser a resposta correta e das outras respostas recebidas para a mesma tarefa não serem. O cálculo dessa probabilidade é formalizado na

Equação 6.1.

$$C(G_a) = \frac{P(G_a \text{ correta}) \prod_{i \neq a} P(G_i \text{ incorreta})}{\prod_{j=1}^{n} P(G_j \text{ incorreta}) + \sum_{j=1}^{n} P(G_j \text{ correta}) \prod_{k \neq j} P(G_k \text{ incorreta})} \tag{6.1}$$

Esse cálculo é inspirado na credibilidade de grupos de respostas em sistemas de computação voluntária (SARMENTA, 2002). O grupo G_a com maior valor de credibilidade c_a é o grupo que representa a resposta candidata da tarefa. Um exemplo da geração de grupos de respostas é apresentado na Figura 6.2.

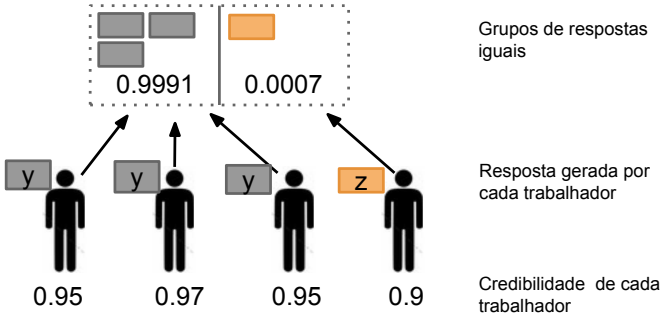


Figura 6.2: Exemplo da geração de um grupos de respostas. Mostra dois grupos de respostas $G_y = \{0.95, 0.97, 0.95\}$ e $G_z = \{0.9\}$. A credibilidade do grupo de respostas y é $C(G_y) = 0.9991$ e a credibilidade do grupo de respostas z é $C(G_z) = 0.0007$.

Algoritmo 3: Pega o Grupo Mais Crível.

Entrada: Mapa G;
Saída: Resposta do grupo mais crível $respGrupo$, Credibilidade do grupo mais crível $credGrupo$;

```
1  respGrupo ← NULL; // Mantém a resposta do grupo de maior credibilidade
2  credGrupo ← 0;    // Mantém a credibilidade do grupo de maior credibilidade
3  para cada chave a em G faça
4      c_a ← C(G, a); /* Calcula a credibilidade do grupo de respostas a pela Eq. 6.1 */
5      se c_a > credGrupo então
6          credGrupo ← c_a;
7          respGrupo ← a;
8      fim
9  fim
10 retorna respGrupo, credGrupo;
```

O laço *repita...até*, no Algoritmo 1, indica que novas réplicas da tarefa são geradas até que pelo menos uma das condições seguintes seja satisfeita:

- Critério de credibilidade é atingido ($credGrupo \geq credRequ$). Isso ocorre quando é obtida a resposta de um grupo que possui um grau de credibilidade igual ou maior que o nível mínimo de credibilidade definido pelo usuário.

- Limite máximo de réplicas é atingido ($contRepl = maxRepl$). Isso ocorre quando o número de réplicas geradas pelo algoritmo atinge o limite máximo de réplicas definido pelo usuário.

Naturalmente, a replicação pode ser concluída por que o número máximo de réplicas (parâmetro $maxRepl$) foi atingido, mas sem que uma resposta final com credibilidade requerida (parâmetro $credRequ$) tenha sido obtida. Nesse caso, o usuário pode avaliar o resultado em uma perspectiva conservadora ou não conservadora. Na perspectiva *conservadora*, o usuário só está interessado nas respostas para as tarefas que atingiram o limite mínimo de credibilidade. As tarefas que não atingiram esse limite são consideradas "sem conclusão". Por outro lado, na perspectiva *não conservadora*, o usuário está interessado na resposta de maior credibilidade, mesmo quando o limite mínimo de credibilidade não pode ser atingido.

Conceitualmente, o algoritmo proposto pode ser definido como de replicação ativa com nível de replicação definido de forma adaptativa. Trata-se de replicação ativa porque cada réplica de uma tarefa é executada integralmente por cada trabalhador, não há *backup* do estado da tarefa durante a execução. A replicação é dita adaptativa no sentido de que o número de réplicas que são geradas para cada tarefa não é pré-definido. Ele é calculado em tempo de execução e é dependente das respostas recebidas dos trabalhadores.

6.3 Materiais e Métodos de Avaliação

Este estudo é baseado em dados obtidos de projetos reais de computação por humanos. Tais dados são descritos na próxima seção. Em seguida, os métodos utilizados na avaliação são detalhados.

6.3.1 Descrição dos Projetos Estudados

O estudo da replicação de tarefas impõe diversos requisitos em termos de base de dados, pois tais bases de dados são utilizadas como rastro na simulação do algoritmo proposto. Todos os

dados descritos para o estudo da credibilidade também são necessários no estudo da replicação. A existência de diversas respostas por tarefa é necessária para que se possa verificar se o algoritmo de replicação obteria uma resposta com menos redundância do que a existente na base de dados. A informação temporal da execução de tarefas é necessária para simular a dinâmica do projeto. Tarefas *ground truth*, aquelas cujas respostas corretas/esperadas sejam conhecidas, são necessárias para que se possa avaliar se o algoritmo de replicação obtém uma resposta correta.

Foram obtidas duas bases de dados que possuem todas as informações necessárias ao estudo de replicação de tarefas: Julgamento de Fatos e Análise de Sentimentos. Essas bases de dados foram apresentadas e utilizadas nos estudos descritos nos capítulos anteriores. No que se refere ao estudo da replicação de tarefas, cabe apenas acrescentar que na base de dados do projeto Julgamento de Fatos há um conjunto de 576 tarefas *ground truth* e que na base de dados do projeto Análise de Sentimentos existem disponíveis 300 tarefas *ground truth*.

6.3.2 Método de Avaliação do Algoritmo de Replicação

A avaliação do algoritmo de replicação é realizada por meio de simulação. A estratégia de replicação proposta foi simulada usando como entrada os dados das tarefas e as respostas providas pelos trabalhadores para cada réplica. A ordem em que as tarefas são executadas e as respostas que os trabalhadores geram são conforme registrados nas bases de dados. Como o algoritmo proposto pode terminar a replicação sem que as respostas de todas as réplicas sejam utilizadas, a ordem em que as respostas armazenadas nas bases de dados são utilizadas tem impacto nos resultados. Esse impacto foi medido por 5 simulações usando as respostas ordenadas de forma aleatória. Na análise dos resultados, sempre se apresenta a média dos resultados obtidos nessas simulações com barras de erros para um nível de confiança de 95%. Naturalmente, o número de réplicas que o algoritmo proposto pode gerar é limitado ao existente na base de dados. Um dos principais objetivos da avaliação é verificar em que situações a estratégia de replicação proposta é capaz de gerar menos réplicas.

Quando a replicação termina porque o número máximo de réplicas foi atingido, mas sem satisfazer o critério de credibilidade, duas situações são possíveis, dependendo dos interesses do usuário. Em uma *perspectiva conservadora*, as tarefas que não atingem o limiar de credibilidade requerida são marcadas como “sem conclusão” e não têm uma resposta associ-

ada. Por outro lado, em uma *perspectiva não conservadora*, a resposta para a tarefa é aquela que obteve o maior valor de credibilidade, mesmo que abaixo do limiar requerido. As duas perspectivas são consideradas na análise dos resultados.

Conforme descrito na Tabela 6.1, todos os parâmetros do algoritmo de replicação (variáveis independentes) são variados nas simulações, são eles: métrica de credibilidade, credibilidade requerida e urgência. Os resultados do efeito da variação dos parâmetros no desempenho do algoritmo foram medidos por meio de três métricas de avaliação (variáveis dependentes), definidas como segue:

- **Economia de réplicas:** É a proporção de economia de réplicas em relação à existente na base de dados. Dado que na base de dados existem x réplicas e a estratégia proposta gerou y réplicas, a economia de réplicas é dada por $\frac{x-y}{x}$. Essa economia é calculada por tarefa e em todo o projeto.

- **Acurácia:** É a taxa de acerto nas tarefas *ground truth*. Ela é calculada como a razão entre o total de tarefas cujas respostas obtidas pelo algoritmo coincidem com as existentes na base de dados *ground truth* e o total de tarefas *ground truth*. Por exemplo, se existem 300 tarefas *ground truth* e para 30 dessas tarefas o algoritmo obteve respostas iguais à existente na base de dados, a acurácia é de 0,1.

- **Tarefas sem conclusão:** É a proporção de tarefas para as quais o algoritmo proposto não atingiu o limiar de credibilidade definido pelo usuário. É calculada como a razão entre o número de tarefas que não atingiram o limiar de credibilidade e o número total de tarefas.

São avaliadas 160 diferentes configurações do algoritmo de replicação proposto (combinações das variáveis independentes, Tabela 6.1). Para identificar as melhores configurações em termos do efeito gerado nas variáveis dependentes, utilizou-se uma análise de otimização multiobjetivo baseada no conceito de fronteira de Pareto (MICHALEWICZ; FOGEL, 2004). A principal ideia dessa otimização é distinguir o conjunto de configurações dominantes e o conjunto de configurações dominadas. As configurações dominantes são aquelas que sempre apresentam resultados superiores aos apresentados pelas configurações dominadas. As configurações dominadas são ditas inviáveis no sentido de que elas não geram melhores resultados

Tabela 6.1: Resumo das variáveis independentes, variáveis dependentes e cenários de referência considerados no estudo da replicação de tarefas.

Variáveis Independentes	
Métrica de credibilidade	Concord. Experimentada, Concord. Reputada, Concord. Presumida, Concord. Simples
Credibilidade requerida	0,6; 0,7; 0,8; 0,91; 0,93; 0,95; 0,97; 0,99
Urgência	0; 0,25; 0,5; 0,75; 1
Variáveis Dependentes	
Economia de réplicas	
Acurácia	
Proporção de tarefas sem conclusão	
Cenários de Referência	
Valor de referência mínimo	Voto majoritário
Valor de referência máximo	Oráculo

em nenhuma variável dependente. Na perspectiva conservadora, tem-se três objetivos: maximizar a economia de réplicas, minimizar a proporção de tarefas sem conclusão e maximizar a acurácia das respostas de tarefas com conclusão. Na perspectiva conservadora, tem-se dois objetivos: maximizar a economia de réplicas e maximizar a acurácia das respostas.

São utilizados dois cenários de referência para avaliação de desempenho do algoritmo: voto majoritário e oráculo. O voto majoritário (valor de referência mínimo) sempre utiliza todas as réplicas existentes na base de dados e elege como resposta final para cada tarefa a resposta provida pela maioria dos trabalhadores que executaram a tarefa. O oráculo (valor de referência máximo) conhece a probabilidade de o trabalhador responder a tarefa corretamente, sendo essa probabilidade 0 ou 1. Quando um trabalhador provê uma resposta correta, ele interrompe a replicação. A economia de réplicas obtida pelo oráculo é a maior economia possível de se obter sem afetar a acurácia. A acurácia obtida por ele é a maior possível. Note, entretanto, que uma estratégia que interrompa a replicação antes que uma resposta correta seja obtida pode economizar mais réplicas do que o oráculo, mas com um custo associado em termos da redução da acurácia.

6.4 Apresentação e Análise dos Resultados

Nesta seção, avalia-se o desempenho do algoritmo de replicação proposto para melhorar o uso que o sistema faz do poder cognitivo disponível por meio da otimização do nível de

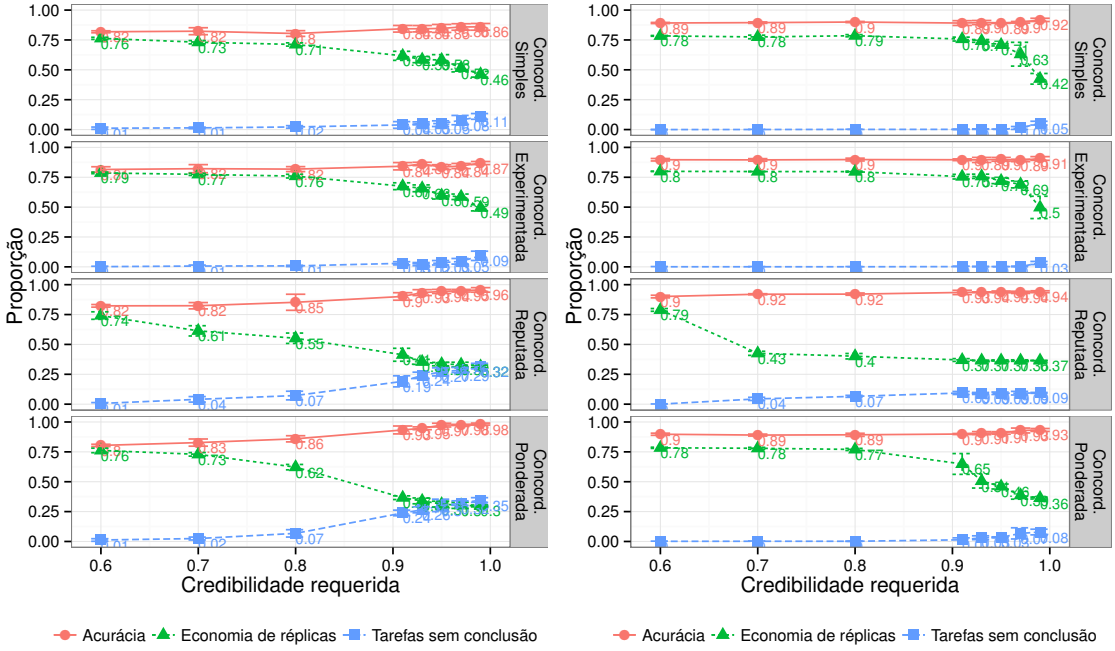
redundância utilizada em cada tarefa. Primeiro, avalia-se em que medida os parâmetros utilizados (métrica de credibilidade, o valor de credibilidade requerida e a urgência) impactam o desempenho do algoritmo em termos da economia de réplicas, acurácia das respostas e tarefas sem conclusão. Após isso, analisa-se como esses parâmetros podem ser escolhidos em uma perspectiva de otimização multiobjetivo das métricas de desempenho. Finalmente, avalia-se como o desempenho do algoritmo se compara à estratégia de voto majoritário com replicação fixa e ao oráculo propostos como cenários de referência.

6.4.1 Configurações e Desempenho

A Figura 6.3 apresenta os resultados da proporção de acurácia, economia de réplicas e tarefas sem conclusão quando se varia o valor de credibilidade requerida nos projetos Análise de Sentimentos e Julgamento de Fatos. Observa-se que *a variação do valor de credibilidade requerida tem grande impacto no desempenho do algoritmo*. Independentemente do projeto e da métrica de credibilidade utilizada, há uma tendência de que quanto maior a credibilidade requerida, menor é a economia de réplicas e maior é o número de tarefas sem conclusão e a acurácia nas tarefas com conclusão.

Observa-se também que há diferenças no desempenho do algoritmo dependendo da métrica de credibilidade utilizada. Comparado com as demais métricas, quando a credibilidade requerida se aproxima de 1, as métricas concordância reputada e concordância ponderada tendem a atingir maiores valores de acurácia, gerarem uma maior proporção de tarefas sem conclusão e menor economia de réplicas. Isso se dá em razão da característica mais conservadora dessas métricas. Elas tendem a gerar valores menores de credibilidade dos trabalhadores comparado aos valores gerados pelas demais métricas. Isso, combinado com uma credibilidade requerida maior, faz com que o algoritmo fique menos propenso a interromper a replicação.

A Figura 6.4, por sua vez, apresenta os resultados da proporção de acurácia, economia de réplicas e tarefas sem conclusão quando se varia o parâmetro de urgência. O comportamento geral que se observa em ambos os projetos e que independe da métrica de credibilidade utilizada é que *aumentar a urgência tende a acarretar uma redução na economia de réplicas e um aumento no número de tarefas sem conclusão e na acurácia das respostas*. Ao analisar esse resultado, deve-se notar que todas as réplicas de cada tarefa são geradas e executadas



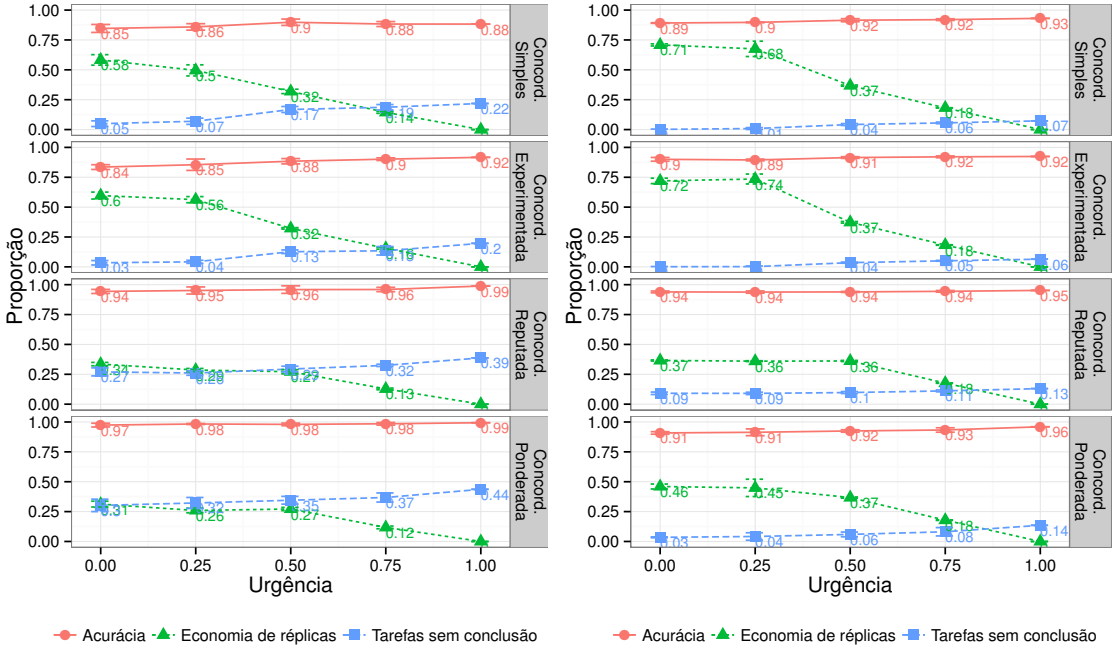
(a) Análise de Sentimentos

(b) Julgamento de Fatos

Figura 6.3: Proporção de acurácia, economia de réplicas e tarefas sem conclusão gerados pelo algoritmo de replicação quando se varia o valor de credibilidade requerida e a métrica de credibilidade utilizada nos projetos (a) Análise de Sentimentos e (b) Julgamento de Fatos. Urgência definida com o valor igual a 0. Mostram-se intervalos para um nível de confiança estatística de 95%.

de uma só vez quando a urgência é definida com o valor igual a 1. Portanto, nesse caso, não há economia de réplica. O benefício de usar o algoritmo de replicação nesse caso é que a escolha da resposta final para a tarefa é baseada na credibilidade dos trabalhadores que proveram as respostas. Por outro lado, quando a urgência é definida com o valor igual a 0, as réplicas de uma tarefa são geradas e executadas uma por vez. Nesse caso, o algoritmo pode detectar quando uma resposta crível foi obtida e interromper a replicação, gerando economia de réplicas, como se observa na Figura 6.4(b).

Além dos parâmetros de credibilidade requerida e de urgência, o desempenho do algoritmo de replicação também depende da dificuldade das tarefas que estão sendo replicadas. A Figura 6.5 mostra a relação entre os graus de dificuldade das tarefas e as credibilidades das respostas obtidas pelo algoritmo. Os resultados apresentados nessa figura indicam que as credibilidades das respostas estão negativamente correlacionadas com o grau de dificuldade das tarefas. Ou seja, *há uma tendência de que quanto mais difícil for a tarefa, menos*



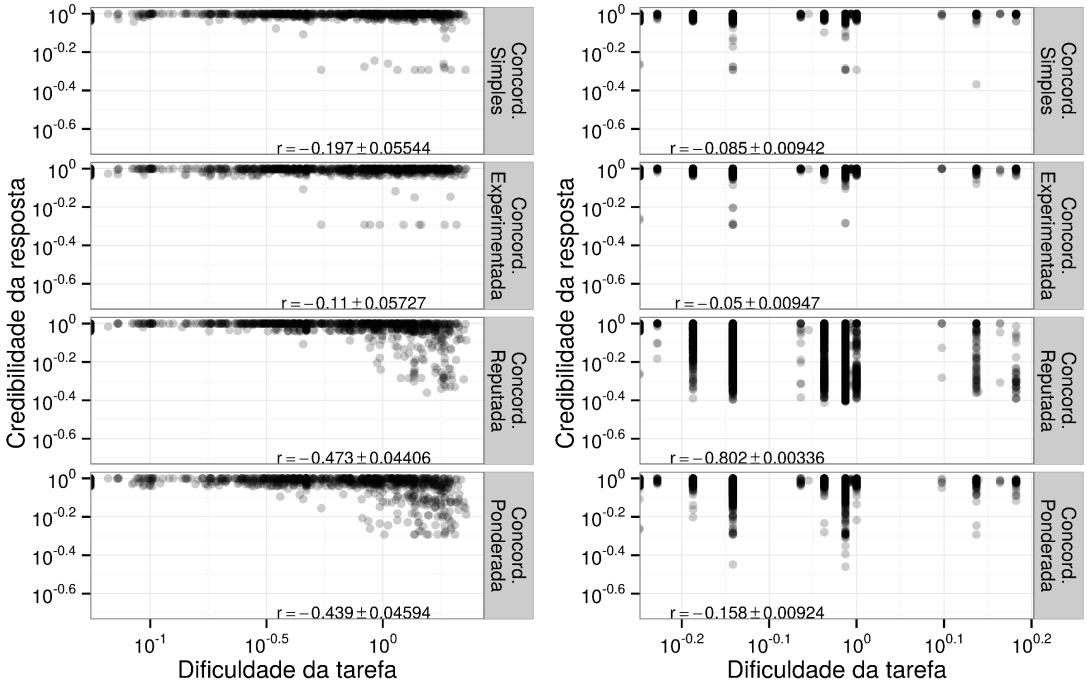
(a) Análise de Sentimentos

(b) Julgamento de Fatos

Figura 6.4: Proporção de acurácia, economia de réplicas e tarefas sem conclusão gerados pelo algoritmo de replicação quando se varia o valor de urgência e a métrica de credibilidade utilizada nos projetos (a) Análise de Sentimentos e (b) Julgamento de Fatos. Credibilidade requerida definida em 0,95. Mostram-se intervalos para um nível de confiança estatística de 95%.

provável é que uma resposta altamente crível seja obtida pelo algoritmo. Essa correlação é maior quando o algoritmo de replicação utiliza as métricas concordância reputada (ρ=-0,51 no projeto Análise de Sentimentos e ρ=-0,8 no projeto Julgamento de Fatos) e concordância ponderada (ρ=-0.44 no projeto Análise de Sentimentos e ρ=-0,16 no projeto Julgamento de Fatos).

Analisou-se também como as economias de réplicas obtidas pelo algoritmo se relacionam com os graus de dificuldade das tarefas nos projetos estudados (Figura 6.6). O resultado obtido nesta análise mostra que a economia de réplicas nas tarefas está negativamente relacionada com o grau de dificuldade das tarefas. Isso indica que *quanto mais difícil são as tarefas menor tende a ser a economia de réplicas gerada pelo algoritmo*. Essa correlação tende a ser maior no projeto Análise de Sentimentos, em que há maior diversidade de dificuldade. No projeto Julgamento de Fatos, a correlação é forte apenas quando a métrica concordância reputada é utilizada (ρ=-0,74).



(a) Análise de Sentimentos (b) Julgamento de Fatos

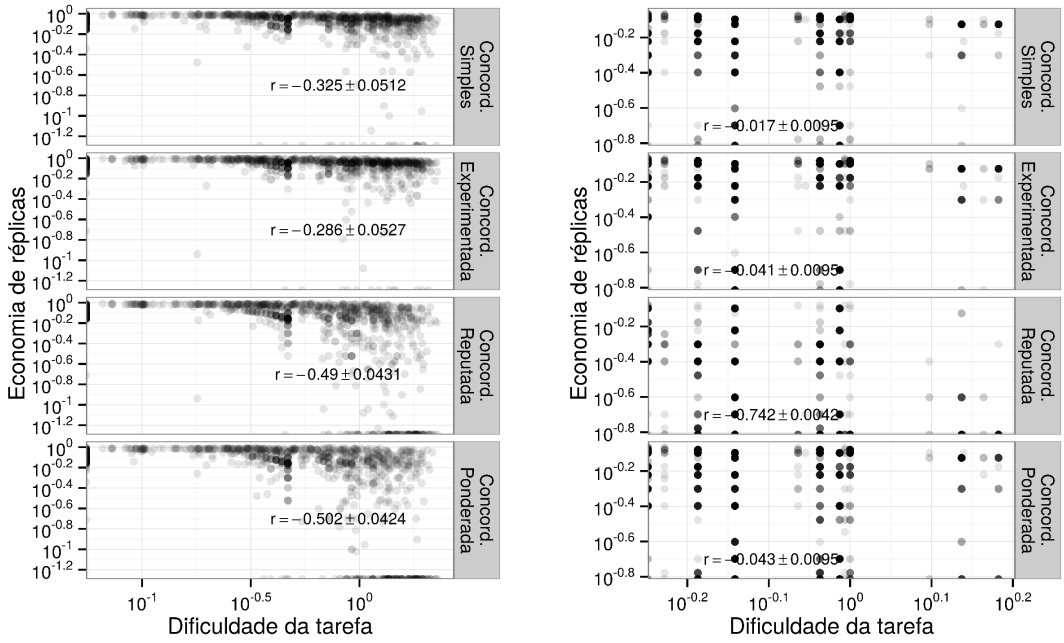
Figura 6.5: Relação entre dificuldade das tarefas e credibilidade da resposta obtida pelo algoritmo de replicação nos projetos (a) Análise de Sentimentos e (b) Julgamento de Fatos. Cada ponto na imagem é uma tarefa. Apresenta a correlação de Spearman e os intervalos de erro para um nível de confiança estatística de 95%.

6.4.2 Análise das Melhores Configurações

Foram testadas 160 diferentes configurações do algoritmo em cada projeto. Tais configurações diferem entre si em termos do valor de urgência, do valor de credibilidade requerida e de métrica de credibilidade usada. Mostra-se importante identificar quais as configurações que apresentam melhor desempenho. Isso pode ser feito considerando a perspectiva conservadora, na qual se admite tarefas sem conclusão, e a perspectiva não conservadora, na qual não se admite tarefas sem conclusão.

Perspectiva Conservadora

Neste caso, tem-se três objetivos a serem otimizados: maximizar a economia de réplicas, minimizar a proporção de tarefas sem conclusão e maximizar a acurácia das respostas de tarefas com conclusão. Existem 13 (8%) configurações dominantes no projeto Julgamento de Fatos (Tabela 6.2) e 26 (16%) configurações dominantes no projeto Análise de Sentimentos



(a) Análise de Sentimentos (b) Julgamento de Fatos

Figura 6.6: Relação entre os graus de dificuldade das tarefas e as economias de réplicas obtidas pelo algoritmo de replicação nos projetos (a) Análise de Sentimentos e (b) Julgamento de Fatos. Cada ponto na imagem é uma tarefa. Apresenta a correlação de Spearman e os intervalos de erro para um nível de confiança estatística de 95%.

(Tabela 6.3). Em ambos os projetos, as configurações diferem entre si em termos dos quatro parâmetros credibilidade requerida, urgência e métrica de credibilidade. É importante notar, entretanto, que *nenhuma configuração inclui a métrica de credibilidade concordância simples*. Esse resultado indica que, comparada às demais métricas de credibilidade, a métrica concordância simples não é a melhor em otimizar nenhuma das três métricas de desempenho avaliadas.

Perspectiva não Conservadora

Em uma perspectiva não conservadora, escolhe-se a resposta de maior credibilidade, quando o algoritmo de replicação atinge o limite de réplicas sem que uma resposta com credibilidade igual ou superior tenha sido obtida. Ou seja, não há tarefas sem conclusão. Nessa análise, tem-se dois objetivos a serem otimizados: maximizar a economia de réplicas e maximizar a acurácia das respostas.

A Figura 6.7 apresenta o desempenho de cada uma das 160 configuração em termos

Tabela 6.2: Configurações dominantes na otimização da economia de réplicas, proporção de tarefas sem conclusão e acurácia em tarefas com conclusão no projeto Julgamento de Fatos.

Configuração do algoritmo			Desempenho do algoritmo		
Cred. Req.	Urgência	Métrica de credibilidade	Economia de réplicas	Acurácia em tarefas com conclusão	Tarefas sem conclusão
0,60	0,00	Experimentada	0,80 ± 0,00	0,90± 0,01	0,00 ±0,00
0,70	0,00	Experimentada	0,80 ± 0,00	0,90± 0,01	0,00 ±0,00
0,93	0,00	Reputada	0,37 ± 0,00	0,94± 0,00	0,09 ±0,00
0,80	0,25	Reputada	0,39 ± 0,02	0,93± 0,00	0,07 ±0,01
0,91	0,25	Ponderada	0,62 ± 0,05	0,91± 0,02	0,02 ±0,01
0,95	0,25	Ponderada	0,43 ± 0,05	0,92± 0,02	0,05 ±0,02
0,97	0,50	Experimentada	0,38 ± 0,00	0,92± 0,01	0,04 ±0,01
0,97	0,50	Reputada	0,36 ± 0,00	0,95± 0,01	0,10 ±0,01
0,99	0,50	Experimentada	0,37 ± 0,00	0,93± 0,01	0,06 ±0,01
0,99	0,50	Reputada	0,36 ± 0,00	0,95± 0,01	0,10 ±0,01
0,80	1,00	Ponderada	0,00 ± 0,00	0,96± 0,00	0,13 ±0,00
0,91	1,00	Ponderada	0,00 ± 0,00	0,96± 0,00	0,13 ±0,00
0,93	1,00	Ponderada	0,00 ± 0,00	0,96± 0,00	0,13 ±0,00

desses objetivos. Esse resultado mostra que diferentes configurações do algoritmo geram diferentes desempenhos em termos da economia de réplicas e acurácia das respostas. Também são exibidos os resultados obtidos pelos cenários de referência, estratégia de replicação fixa com voto majoritário e oráculo.

Considerando ambos os objetivos de maximização da acurácia e da economia de réplicas, obteve-se o conjunto de configurações dominantes no conjunto de configurações das 160 configurações testadas. Trata-se de configurações que são melhores que as demais em ambos os objetivos. Obteve-se 6 (4%) configurações dominantes no projeto Análise de Sentimentos (Tabela 6.4) e 3 (2%) configurações dominantes no projeto Julgamento de Fatos (Tabela 6.5).

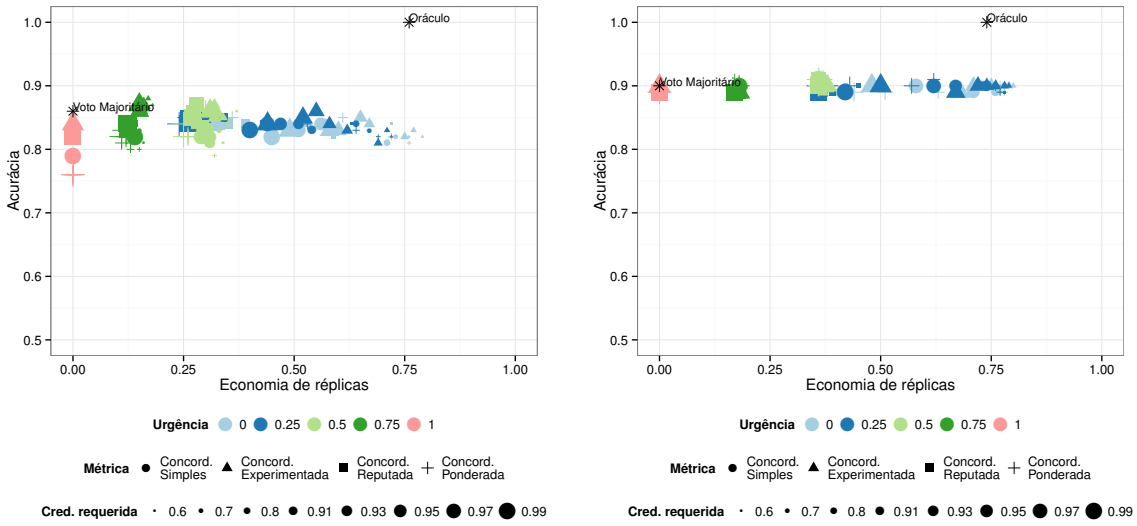
A escolha de uma dessas configurações depende do interesse do usuário. Em algumas situações, pode-se dar maior prioridade à acurácia das respostas, à economia de réplicas, ou a ambas. Observa-se que *apenas as métricas de credibilidade concordância experimentada e concordância reputada aparecem entre as melhores configurações*. Portanto, são métricas que melhor otimizam a acurácia e economia de réplicas. Isso é um indicativo de que, em termos dos objetivos de interesse, é importante levar em conta a aleatoriedade na concordância entre os trabalhadores (como faz a concordância experimentada) e histórico de credibilidade daqueles com os quais a concordância se dá (como faz a concordância reputada). No caso do

Tabela 6.3: Configurações dominantes na otimização da economia de réplicas, proporção de tarefas sem conclusão e acurácia em tarefas com conclusão no projeto Análise de Sentimentos.

Configuração do algoritmo			Desempenho do algoritmo		
Cred. Req.	Urgência	Métrica de credibilidade	Economia de réplicas	Acurácia em tarefas com conclusão	Tarefas sem conclusão
0,60	0,00	Experimentada	0,78 ± 0,01	0,83 ± 0,02	0,00 ± 0,00
0,60	0,00	Reputada	0,72 ± 0,02	0,84 ± 0,02	0,01 ± 0,00
0,70	0,00	Experimentada	0,77 ± 0,01	0,84 ± 0,04	0,01 ± 0,00
0,80	0,00	Ponderada	0,61 ± 0,01	0,87 ± 0,02	0,08 ± 0,02
0,80	0,00	Reputada	0,53 ± 0,03	0,88 ± 0,04	0,09 ± 0,02
0,91	0,00	Ponderada	0,36 ± 0,01	0,95 ± 0,02	0,24 ± 0,02
0,91	0,00	Experimentada	0,66 ± 0,03	0,85 ± 0,02	0,03 ± 0,01
0,91	0,00	Reputada	0,39 ± 0,03	0,91 ± 0,02	0,21 ± 0,03
0,93	0,00	Ponderada	0,33 ± 0,01	0,96 ± 0,01	0,28 ± 0,02
0,93	0,00	Experimentada	0,64 ± 0,02	0,86 ± 0,03	0,02 ± 0,01
0,95	0,00	Ponderada	0,31 ± 0,02	0,97 ± 0,01	0,31 ± 0,03
0,97	0,00	Ponderada	0,30 ± 0,01	0,98 ± 0,01	0,33 ± 0,01
0,99	0,00	Ponderada	0,29 ± 0,01	0,99 ± 0,01	0,36 ± 0,01
0,91	0,25	Ponderada	0,29 ± 0,02	0,96 ± 0,01	0,27 ± 0,02
0,91	0,25	Reputada	0,33 ± 0,03	0,93 ± 0,02	0,22 ± 0,02
0,95	0,25	Experimentada	0,55 ± 0,02	0,87 ± 0,03	0,04 ± 0,01
0,60	0,50	Experimentada	0,37 ± 0,00	0,88 ± 0,02	0,05 ± 0,01
0,80	0,50	Experimentada	0,34 ± 0,00	0,89 ± 0,01	0,10 ± 0,02
0,80	0,50	Reputada	0,30 ± 0,01	0,94 ± 0,01	0,23 ± 0,01
0,91	0,50	Experimentada	0,32 ± 0,01	0,91 ± 0,02	0,12 ± 0,02
0,91	0,50	Reputada	0,28 ± 0,01	0,96 ± 0,01	0,26 ± 0,02
0,97	0,50	Reputada	0,28 ± 0,01	0,98 ± 0,01	0,30 ± 0,02
0,60	0,75	Experimentada	0,18 ± 0,01	0,89 ± 0,02	0,07 ± 0,02
0,70	0,75	Experimentada	0,17 ± 0,01	0,91 ± 0,01	0,08 ± 0,01
0,99	0,75	Ponderada	0,11 ± 0,01	1,00 ± 0,01	0,41 ± 0,02
0,99	0,75	Experimentada	0,15 ± 0,00	0,92 ± 0,02	0,19 ± 0,02

projeto Julgamento de Fatos, observa-se um empate entre as duas configurações com maior economia de réplicas e menor acurácia. Quando isso acontece, ambas as configurações são mantidas como configurações dominantes. A escolha de uma delas pode depender de outro aspecto a ser considerado pelo usuário que não a acurácia ou a economia de réplicas.

Observa-se também que *algumas configurações apresentam maior acurácia e economia de réplicas que a estratégia de replicação fixa com voto majoritário*. Esse resultado mostra que o fato de o algoritmo obter um número menor de respostas por tarefas e escolher a resposta final a partir da credibilidade dos trabalhadores não implica em ele obter uma proporção de respostas erradas maior do que a estratégia de voto majoritário. Pelo contrário,



(a) Análise de Sentimentos

(b) Julgamento de Fatos

Figura 6.7: Proporção de economia de réplicas e de acurácia obtidas por diferentes configurações do algoritmo de replicação. Apresentam-se os dois cenários de referência: replicação fixa com voto majoritário e oráculo.

há situações em que a acurácia obtida pelo algoritmo de replicação proposto é maior que a acurácia obtida pelo voto majoritário. Naturalmente, se o usuário definir um nível de credibilidade requerida muito baixo ou uma urgência muito alta, a acurácia e a economia de réplicas serão comprometidos. Esse é o caso das configurações não dominantes.

Finalmente, *algumas configurações do algoritmo proposto superam o oráculo em termos de economia de réplicas e aproxima-se dele em termos de acurácia*. Note que a economia de réplicas obtida pelo oráculo é a maior economia possível de se obter sem afetar a acurácia. A acurácia obtida por ele é a maior possível. Não se sabe se é possível obter uma acurácia perfeita como a gerada pelo oráculo. Em computação por humanos é comum existir um conjunto de tarefas em que a divergência entre os trabalhadores é tamanha que apenas um especialista é capaz de decidir sobre qual resposta é mais adequada (SHENG; PROVOST; IPEIROTIS, 2008; AROYO; WELTY, 2014). Mesmo métodos estatísticos dedicados a aumentar a acurácia não atingem acurácia perfeita como a gerada pelo oráculo (SHESHADRI; LEASE, 2013). Em vista disso, é satisfatório o fato dos resultados obtidos pela estratégia de replicação proposta, cujo principal objetivo é economizar réplicas, serem melhores que os resultados obtidos pela estratégia de voto majoritário quando considera-se o número de réplicas e a acurácia e próximos aos resultados obtidos pelo oráculo quando considera-se a acurácia.

Tabela 6.4: Configurações dominantes na otimização da economia de réplicas e acurácia total no projeto Análise de Sentimentos.

Configuração do algoritmo			**Desempenho do algoritmo**	
Cred. Req.	Urgência	Métrica de credibilidade	Economia de réplicas	Acurácia total
0,60	0,00	Experimentada	$0{,}78 \pm 0{,}01$	$0.83 \pm 0{,}02$
0,70	0,00	Experimentada	$0{,}77 \pm 0{,}01$	$0.84 \pm 0{,}04$
0,91	0,00	Experimentada	$0{,}66 \pm 0{,}03$	$0.85 \pm 0{,}03$
0,93	0,00	Experimentada	$0{,}64 \pm 0{,}03$	$0.86 \pm 0{,}01$
0,91	0,25	Reputada	$0{,}33 \pm 0{,}03$	$0.87 \pm 0{,}04$
0,70	0,75	Experimentada	$0{,}17 \pm 0{,}01$	$0.89 \pm 0{,}02$
Cenários de referência				
Replicação Fixa			0,00	0,86
Oráculo			$0{,}76 \pm 0{,}01$	1,00

Tabela 6.5: Configurações dominantes na otimização da economia de réplicas e acurácia total no projeto Julgamento de Fatos.

Configuração do algoritmo			**Desempenho do algoritmo**	
Cred. Req.	Urgência	Métrica de credibilidade	Economia de réplicas	Acurácia total
0,60	0,00	Experimentada	$0{,}80 \pm 0{,}00$	$0{,}90 \pm 0{,}01$
0,70	0,00	Experimentada	$0{,}80 \pm 0{,}00$	$0{,}90 \pm 0{,}01$
0,60	0,00	Reputada	$0{,}78 \pm 0{,}01$	$0{,}91 \pm 0{,}01$
Cenários de referência				
Replicação Fixa			0,00	0,90
Oráculo			$0{,}74 \pm 0{,}01$	1,00

6.5 Considerações Finais

Neste capítulo, discutiu-se o mecanismo de replicação de tarefas e como esse mecanismo pode ser utilizado a fim de melhorar o uso do poder cognitivo disponível em projetos de computação por humanos. Propôs-se um algoritmo de replicação com esse objetivo. O algoritmo proposto visa otimizar a replicação de tarefas de modo a tratar compromissos entre economia de réplicas, acurácia das respostas e tarefas sem conclusão. O algoritmo permite que os usuários definam o nível de urgência desejada, o nível de credibilidade requerida nas respostas, e a métrica de credibilidade a ser utilizada.

Avaliou-se o algoritmo usando dados de 2 projetos de computação por humanos reais. O desempenho do algoritmo é medido pela proporção de economia de réplicas, proporção de acurácia e proporção de tarefas sem conclusão. O desempenho exibido pelo algoritmo proposto é comparado com o exibido por uma estratégia de referência inferior (voto majori-

tário) e uma estratégia de referência superior (oráculo). No geral, os resultados da avaliação revelam que:

- O desempenho do algoritmo é fortemente dependente dos parâmetros utilizados: (i) quanto maior a credibilidade requerida pelo usuário, menor é a economia de réplicas e maior é o número de tarefas sem conclusão e maior a acurácia nas tarefas com conclusão; (ii) quanto maior a urgência requerida pelo usuário, menor a economia de réplicas, maior a proporção de tarefas sem conclusão e maior a acurácia das respostas em tarefas com conclusão;

- O desempenho do algoritmo é fortemente dependente das características das tarefas que estão sendo replicadas: quanto mais difícil for a tarefa, menor tende a ser a economia de réplicas e a credibilidade das respostas obtidas pelo algoritmo;

- As melhores configurações do algoritmo proposto superam o oráculo em termos de economia de réplicas e superam a estratégia de voto majoritário em termos de acurácia das respostas;

- Quando não se admite tarefas sem conclusão, as métricas de credibilidade concordância experimentada e concordância reputada apresentam os melhores desempenhos de otimização da economia de réplicas e da acurácia;

- Quando se admite tarefas sem conclusão, apenas a métrica de credibilidade concordância simples não compõe as melhores configurações de otimização dos objetivos.

O estudo apresentado neste capítulo mostra que o conceito de replicação adaptativa se coloca de forma satisfatória no contexto de projetos de computação por humanos. O algoritmo de replicação proposto permite otimizar o uso do poder cognitivo dos trabalhadores em projetos de computação por humanos ao gerar um número de réplicas não maior que o suficiente para que uma resposta crível seja obtida.

Capítulo 7

Limitações

O propósito deste capítulo é discutir as limitações das métricas e algoritmos propostos e da avaliação realizada nos capítulos anteriores. Há dois tipos principais de limitações que precisam ser destacados. O primeiro se refere às restrições de uso das métricas e algoritmos propostos. O segundo tipo de limitação se refere às ameaças à validade da avaliação realizada e dos resultados obtidos. Essas limitações são discutidas nas seções seguintes.

7.1 Restrições das Métricas e Algoritmos Propostos

Mostra-se relevante explicitar o espaço de aplicação das métricas e algoritmos propostos no contexto de computação por humanos. Trata-se de características da forma como as métricas e os algoritmos são construídos e que restringem os contextos aos quais eles se limitam. Essa discussão é apresentada nesta seção pela análise dos tipos de tarefa, tipos de aplicação e tipos de sistema de computação por humanos.

Quanto ao tipo de tarefa de computação por humanos, as métricas propostas podem ser utilizadas tanto em tarefas factuais quanto em tarefas não factuais. A avaliação reportada neste documento se ateve a tarefas factuais em razão da limitação das bases de dados disponíveis. Entretanto, não há nada na forma como as métricas são construídas que as restrinja às tarefas classificadas como factuais.

As métricas de credibilidade e de dificuldade são baseadas na concordância existente no conjunto de respostas recebido para cada tarefa. Isso permite que essas métricas não façam qualquer pre-consideração sobre o tipo de dado de entrada e das instruções da tarefa. Assim,

elas podem ser utilizadas em tarefas cujos dados de entrada são os mais diversos, como: imagens, mensagens de texto, áudio, vídeos, etc. No entanto, as métricas possuem restrição quanto ao formato das respostas geradas pelos trabalhadores para as tarefas. As métricas assumem que as respostas providas pelos trabalhadores sejam estruturadas de modo que possam ser agregadas para se detectar convergências e divergências. Tarefas com essas características são comuns em computação por humanos, mas também há tarefas cujas respostas não são estruturadas, nas quais tais métricas não se aplicariam. Esse é o caso, por exemplo, de tarefas nas quais as respostas geradas pelos trabalhadores são textos não estruturados, como uma redação, ou conteúdos visuais, como uma logomarca.

Quanto ao tipo de aplicação de computação por humanos, as métricas de engajamento podem ser usadas tanto em aplicações do tipo projeto quanto em aplicações do tipo *workflow*. Em razão da limitação de bases dados disponíveis, ateve-se apenas a aplicações do tipo projeto. As métricas de credibilidade e o algoritmo de replicação, por sua vez, foram propostos para o escopo específico de aplicações do tipo projeto. Elas pressupõem que os trabalhadores sempre executam os mesmos tipos de tarefas e que tais tarefas diferem entre si apenas quanto ao dado de entrada. Esse é tipo mais comum de aplicação de computação por humanos. É o caso, por exemplo, de projetos como o Galaxy Zoo em que os trabalhadores sempre executam o mesmo tipo de tarefa, mudando-se apenas a imagem da galáxia a que a tarefa se refere. Esse também é o caso de grupos de tarefas que requerem inteligência humana (HITs, do inglês *Human Intelligence Tasks*) no Mturk, em que tipicamente as todas as tarefas que compõem um grupo de HITs são iguais, diferindo apenas em termos do item de entrada.

Quanto ao tipo de sistema de computação por humanos, as métricas propostas e o algoritmo proposto não estão acoplados a nenhum tipo específico de sistema de computação por humanos. A rigor, tais métricas e algoritmo podem ser utilizados na análise do engajamento e da credibilidade e na otimização da replicação de tarefas em projetos executados em sistemas de pensamento voluntário e em mercado de trabalho *online*. É importante destacar que, neste trabalho, cada projeto foi estudado isoladamente sem tratar o sistema como um todo. Naturalmente, quando se considera um sistema com múltiplos projetos, surgem novas questões a serem consideradas. Questões essas que vão além do escopo proposto na pesquisa reportada neste documento. Como forma de exemplificar e fomentar esse tipo estudo,

algumas dessas questões são discutidas no Apêndice C.

A pesquisa conduzida neste trabalho também pode ser analisada sob a perspectiva de outras áreas de pesquisa, principalmente as áreas denominadas *ciência cidadã* e *crowdsourcing*. Todos os projetos estudados na pesquisa descrita neste documento podem ser classificados como projetos de *crowdsourcing*. Quatro deles se inserem no contexto de ciência cidadã, são eles: Galaxy Zoo, The Milky Way Project, Cell Spotting e Sun4All. Assim, é natural se questionar em que medida os resultados de engajamento, credibilidade e replicação obtidos com dados desses projetos podem se generalizados ao contexto de ciência cidadã e de *crowdsourcing*. Ao fazer esse tipo de generalização é importante considerar o fato de que as atividades desempenhadas pelas pessoas em ciência cidadã e em *crowdsourcing* podem ser as mais diversas e não consistirem necessariamente em uma atividade que se baseia nas capacidades cognitivas das pessoas (QUINN; BEDERSON, 2011; WIGGINS; CROWSTON, 2012; LINTOTT; REED, 2013). Considera-se que a aplicação dos resultados reportados neste documento se restringe às iniciativas de ciência cidadã e *crowdsourcing* que sejam baseadas em tarefas de computação por humanos.

7.2 Ameaças à Validade

Ameaças à validade estão em grande parte relacionadas ao tipo de pesquisa conduzida. Quanto à forma de abordagem, a pesquisa descrita neste documento é classificada como quantitativa por focar em aspectos comportamentais mensuráveis. Nenhuma ênfase é dada aos fatores qualitativos do engajamento, credibilidade e dificuldade. Quanto ao objetivo, por sua vez, a pesquisa descrita neste documento combina dois tipos principais de pesquisa científica: pesquisa descritiva e pesquisa explicativa. A pesquisa é descritiva no sentido de que ela elicita as características de engajamento e credibilidade dos trabalhadores em dados obtidos de sistemas reais. A pesquisa é explicativa quando se analisa o impacto que diferentes configurações do algoritmo de replicação de tarefas exercem sobre o desempenho dele. No contexto desses tipos de pesquisas, têm-se três tipos de ameaças à validade que precisam ser discutidas: validade de construção, validade interna e validade externa (CHRISTENSEN, 2007).

A *validade de construção* reflete a extensão em que o estudo feito realmente mede o que

ele se propõe a medir. Nesse contexto, cabe ressaltar que o presente trabalho não se trata do teste de uma teoria, embora teorias sobre engajamento e credibilidade tenham orientado a proposta das métricas utilizadas no estudo. A rigor, o que se faz neste trabalho é articular e reusar conceitos de modo a inspirar a proposta das métricas.

Um ameaça de construção que cabe ser destacada é quanto à métrica de dificuldade das tarefas. Considerando a literatura, utilizou-se a Entropia de Shannon para medir dificuldade percebida pelos trabalhadores (GREITZER, 2005; ASLAM; PAVLU, 2007; ARCANJO et al., 2014; AROYO; WELTY, 2014). Ou seja, assume-se que tarefas em que há maior variação de respostas tendem a ser mais difíceis para os trabalhadores. Entretanto, em situações em que isso não for verdade, essa construção deve ser vista apenas como uma estimativa de incerteza e não de dificuldade. Qualquer métrica quantitativa de dificuldade está sujeita a essa ameaça. Isso não tem implicações no estudo quantitativo de replicação de tarefas, mas sim na forma como se interpreta as relações entre engajamento e dificuldade, e entre credibilidade e dificuldade.

A *validade interna*, por sua vez, reflete em que extensão é possível estabelecer uma conclusão causal com base no estudo reportado neste documento. Nesse contexto, a caracterização do engajamento e da credibilidade dos trabalhadores consistiu apenas na observação de variáveis de interesse sem qualquer manipulação. Apenas identificaram-se as variações no engajamento e na credibilidade, sem se distinguir as causas de tais variações. No estudo das relações entre métricas, realizado por meio de análises de correlação e regressão, não se estabelece uma relação de causa e efeito entre as variáveis estudadas. Além disso, como todo estudo dessa natureza, não se sabe se há outras variáveis não estudadas que podem afetar as correlações e regressões obtidas.

No estudo da replicação de tarefas, manipularam-se as variáveis independentes (urgência, credibilidade requerida, métrica de credibilidade) para verificar os efeitos delas sobre as variáveis dependentes (economia de réplicas, acurácia e tarefas sem conclusão). Nesse estudo, a validade interna da relação estabelecida entre as variáveis se dá nos limites do nível de significância estatística utilizado.

Por fim, a *validade externa* reflete a extensão em que os resultados obtidos podem ser generalizados para outros contextos. O fato das análises reportadas neste documento terem sido realizadas em diversos projetos e de diversos resultados serem semelhantes reforça a validade externa. Entretanto, 6 projetos ainda é um número de projetos que pode ser consi-

derado pequeno. Além disso, por limitações dos dados disponíveis nos projetos, em algumas análises apenas dados de 2 projetos puderam ser utilizados.

Dessa forma, os resultados obtidos neste estudo podem ser considerados uma evidência, mas não uma garantia de que os mesmos comportamentos serão observados em outros projetos. Dados de outros projetos precisam ser analisados para que invariantes no comportamento de engajamento e credibilidade dos trabalhadores sejam identificadas. A partir da observação sistemática desses comportamentos, pode-se propor teorias que delimitem os contextos em que eles ocorrem.

Capítulo 8

Trabalhos Relacionados

Até onde se sabe, a pesquisa descrita neste documento é uma das primeiras (i) a analisar oferta de poder cognitivo de trabalhadores em sistemas de computação por humanos de acordo com a literatura de engajamento de seres humanos (O'BRIEN; TOMS, 2008; RODDEN; HUTCHINSON; FU, 2010) e de acordo com a literatura de credibilidade de seres humanos (WATHEN; BUREL, 2002; RIEH; DANIELSON, 2007) e (ii) a tratar o conceito de replicação de tarefas em sistemas de computação por humanos. Estudos no estado da arte muitas vezes não se referem a essa literatura e nem aos termos engajamento e credibilidade, mas podem ser discutidos por meio dessas lentes. Os estudos no contexto do engajamento são discutidos na Seção 8.1, os estudos no contexto de credibilidade são discutidos na Seção 8.2 e os estudos sobre relações entre engajamento e credibilidade são apresentados na Seção 8.3. Finalmente, estudos correlatos no contexto de replicação de tarefas são analisados na Seção 8.4.

8.1 Engajamento

Muitos dos estudos que até então tinham sido conduzidos com o propósito de entender o comportamento de seres humanos em sistemas de computação por humanos focam principalmente em delinear os fatores psicológicos que explicam o engajamento no sistema. Um dos primeiros estudos a avançar o conhecimento sobre tais fatores foi conduzido por Raddick et al. (2008). Em um estudo qualitativo no projeto Galaxy Zoo, eles mostram que, entre 12 categorias de motivação mencionadas pelos trabalhadores que executam tarefas nesse projeto, a categoria que é mais mencionada é o interesse em astronomia, que é o tema do

projeto. Dessa forma, no caso desse projeto, esse é o principal fator que motiva o início do engajamento cognitivo dos trabalhadores.

Em um estudo qualitativo conduzido no projeto Biotracker, Rotman et al. (2012) destacam que, além de compreender os motivos do início do engajamento, é importante identificar o que mantém os trabalhadores engajados ao longo do tempo e o que causa o ponto de desengajamento. Ou seja, os fatores que determinam o engajamento duradouro são importantes. Os resultados obtidos no estudo mostraram que os trabalhadores se engajam no projeto por interesses pessoais, como curiosidade. No entanto, eles se mantêm engajados executando tarefas no projeto principalmente pelo reconhecimento recebido pelas suas contribuições. A atração de pessoas apenas curiosas sobre o projeto e que não têm interesse em engajamento duradouro pode ser relacionado ao uso de divulgação em massa, por exemplo, em redes sociais (ROBSON et al., 2013). Os trabalhadores que ficam engajados por mais tempo podem exibir um padrão de atuação mais instável ou mais comprometido (EVELEIGH et al., 2014).

Notadamente, esses estudos qualitativos apresentam diversos fatores psicológicos que explicam a dinâmica do processo de engajamento dos trabalhadores em sistemas de computação por humanos. A pesquisa reportada neste documento trouxe para a área de computação por humanos a preocupação com uma medição mais sistemática do engajamento de trabalhadores nos projetos (PONCIANO et al., 2014b; PONCIANO; BRASILEIRO, 2014). Até então não se discutia, por exemplo, o problema de que a maioria dos trabalhadores deixa o projeto após o primeiro dia de participação (transientes). Pouco se podia dizer sobre as características da duração do período de engajamento dos trabalhadores, tempo dedicado diariamente e a variabilidade nos retornos dentro desse período. Também era desconhecido em que medida os trabalhadores diferem entre si em termos dessas características.

Recentemente, surgiram outros estudos que também visam explorar características do engajamento dos trabalhadores seguindo uma abordagem quantitativa. Por exemplo, alguns estudos têm focado em entender em que medida o engajamento de trabalhadores em sistemas de computação por humanos que utiliza a contribuição de voluntários se compara ao engajamento em sistemas em que se paga pela execução das tarefas (SAUERMANN; FRANZONI, 2015). Isso permite estimar monetariamente o valor da contribuição dessas pessoas aos projetos. Outra perspectiva de interesse é visualizar melhor a dinâmica do engajamento dos trabalhadores nos projetos (MORAIS; SANTOS; RADDICK, 2015). Técnicas adequadas de visu-

alização permitem colocar em perspectiva padrões comportamentais que são, muitas vezes, difíceis de serem observados sem o uso de ferramentas apropriadas. No geral, as métricas de engajamento propostas na pesquisa descrita neste documento se assemelham às utilizadas nesses estudos e os resultados acerca do engajamento são congruentes.

Uma linha de estudo que tem sido pouco tratada na literatura e que ganha perspectiva após os estudos quantitativos apresentados neste documento é a associação entre os motivos de engajamento reportados pelos trabalhadores em estudos qualitativos (e.g. Raddick et al. (2008) e Eveleigh et al. (2014)) e o engajamento real que se mede no histórico de tarefas. Uma vez que os resultados obtidos neste trabalho mostram que os trabalhadores apresentam diferentes perfis de engajamento, estudos sobre fatores motivacionais podem ser realizados considerando as peculiaridades de engajamento de cada perfil. Isso permite verificar fatores mais específicos da motivação dos trabalhadores. Por exemplo, os trabalhadores persistentes são mais extrinsecamente motivados do que os trabalhadores que exibem os outros perfis? Um primeiro esforço nesse sentido é a pesquisa reportada por (BEIRNE; LAMBIN, 2013), que relaciona fatores comportamentais de contribuição e dados demográficos e de gênero.

Além de complementar o entendimento do engajamento dos trabalhadores, estudos nessa direção podem fornecer informações que permitam que os projetistas e os gestores de projetos de computação por humanos desenvolvam estratégias que visem otimizá-los (PONCIANO et al., 2014). Por exemplo, estratégias mais focadas em motivar trabalhadores que exibam o perfil de engajamento que ele deseja atrair e reter no projeto. Trata-se de intervenções mais personalizadas (à luz da literatura sobre comportamento humano e métricas de interesse (FISCHER, 2001)). Na pesquisa descrita neste documento, ao prover um conjunto de métricas para medir o engajamento e ao mostrar o engajamento típico dos trabalhadores em diversos projetos, espera-se motivar mais estudos nessa direção.

8.2 Credibilidade

A maior parte dos estudos que tratam da credibilidade de resultados obtidos em computação por humanos tem se concentrado em permitir que uma resposta adequada seja obtida em um conjunto de respostas redundantes (SHESHADRI; LEASE, 2013; BIRD et al., 2014). Essa corrente de estudos foca apenas na agregação de respostas que, via de regra, é um procedimento

offline que é realizado após todas as respostas redundantes serem obtidas do sistema. Trata-se do uso de arcabouços estatísticos que não dão ênfase à credibilidade de cada trabalhador. Elege-se uma resposta final para a tarefa, mas não se discute qual a credibilidade dessa resposta. Essa é a abordagem, por exemplo, do voto majoritário que é uma estratégia muito simples, mas amplamente utilizada na prática (SHESHADRI; LEASE, 2013). Existem diversos outros estudos que seguem essa abordagem, por exemplo, Whitehill et al. (2009) e Dalvi et al. (2013).

A pesquisa descrita neste documento parte da estimativa da credibilidade de cada trabalhador para que se possa agregar respostas providas por diferentes trabalhadores, obter a resposta final para a tarefa e obter uma estimava da credibilidade dessa resposta. São propostas quatro métricas diferentes de medir credibilidade inspiradas nos conceitos definidos em arcabouços de estudo de credibilidade (WATHEN; BUREL, 2002; RIEH; DANIELSON, 2007). Estudos conduzidos em sistemas de computação por humanos que tratam de fatores psicológicos que explicam características de credibilidade dos trabalhadores foram relevantes na definição dessas métricas. Tais estudos diferem quanto ao tipo de sistema de computação por humanos que é abordado: mercados de computação por humanos e sistemas de pensamento voluntário.

Em mercados de computação por humanos, alguns trabalhadores estão mais interessados em receber a remuneração oferecida pela execução, independentemente se a tarefa está sendo executada adequadamente ou não (QUINN; BEDERSON, 2011; KITTUR et al., 2013). Esse comportamento geralmente está associado a problemas no projeto da tarefa (KOCHHAR; MAZZOCCHI; PARITOSH, 2010; EICKHOFF; VRIES, 2011; KAZAI; KAMPS; MILIC-FRAYLING, 2013). Isso significa que tarefas mal projetadas favorecem o surgimento desse tipo de comportamento. Trabalhadores também podem cometer erros intencionais como forma de reagir a ações inadequadas dos usuários. Por exemplo, eles podem planejar um conluio contra os usuários que submetem tarefas mal projetadas (KULKARNI; CAN; HARTMANN, 2012). Em sistemas de pensamento voluntário, a ocorrência desse tipo de comportamento é considerada insignificante (LINTOTT et al., 2008). Mesmo os trabalhadores que apresentam um engajamento instável e que estão menos motivados se preocupam com a qualidade do trabalho que realizam (EVELEIGH et al., 2014).

Independente de tipo de sistema, a ocorrência de respostas inadequadas tende a ser baixa.

Em grande parte das tarefas, os trabalhadores convergem unanimemente para uma mesma resposta. Isso é observado nos projetos estudados na pesquisa descrita neste documento em que há unanimidade na resposta provida pelos trabalhadores em mais de 70% das tarefas do projeto Julgamento de Fatos e em 20% das tarefas no projeto Análise de Sentimentos. Isso também se verifica nas altas credibilidades dos trabalhadores e nas acurácias das respostas. As observações na literatura de que a ocorrência de respostas inadequadas deve-se em grande parte a características da tarefa também são constatados na pesquisa descrita neste documento. Isso é evidente quando se observa que as credibilidades dos trabalhadores e as acurácias das respostas são menores em tarefas difíceis.

Finalmente, enquanto é comum a medição da credibilidade e uso dessa mediação na dinâmica dos sistemas como buscadores Web (RIEH; DANIELSON, 2007; SCHWARZ; MORRIS, 2011) e computação voluntária (SARMENTA, 2002), até onde se sabe isso não tem sido feito em computação por humanos. A pesquisa reportada neste documento buscou um primeiro esforço nesse sentido ao medir a credibilidade dos trabalhadores considerando diferentes aspectos do arcabouço teórico e ao utilizar essas medições na otimização da replicação de tarefas (PONCIANO; BRASILEIRO; GADELHA, 2013; PONCIANO et al., 2014a).

8.3 Relações

Os resultados reportados neste documento mostram que, no contexto de computação por humanos, a dificuldade desempenha um importante papel no estudo do engajamento e da credibilidade dos trabalhadores e na replicação de tarefas. Até onde se sabe, nenhum outro estudo investigou a variação da credibilidade de seres humanos com a dificuldade percebida por ele. Isso ocorre por que nas áreas em que o estudo de credibilidade é mais comum (como, por exemplo: *blogs* (JUFFINGER; GRANITZER; LEX, 2009), *microblogs* (MORRIS et al., 2012), sítios de notícias (ZHANG et al., 2011) e páginas Web em geral (SCHWARZ; MORRIS, 2011)) o conceito de dificuldade não é tratado como relevante.

Os estudos na literatura sobre engajamento de seres humanos, por sua vez, tratam da dificuldade percebida por seres humanos, mas em uma perspectiva diferente. Geralmente a dificuldade da tarefa é vista como um desafio que pode, inclusive, motivar um maior engajamento. Isso ocorre, por exemplo, no contexto de jogos (O'BRIEN; TOMS, 2008). Como reve-

lam os resultados reportados neste documento e em outros existentes na literatura (e.g. Eveleigh et al. (2014), Ipeirotis e Gabrilovich (2014) e Martin et al. (2014)), em computação por humanos, quando o trabalhador percebe dificuldade e acredita que pode estar provendo respostas inadequadas, ele tende a se desengajar. Isso ocorre por temor de atrapalhar uma pesquisa científica ou por temor de ser penalizado pelo usuário.

Quanto à relação entre métricas de engajamento e métricas de credibilidade, o estudo conduzido neste trabalho reforça e complementa esforços de trabalhos anteriores (RZESZOTARSKI; KITTUR, 2011; IPEIROTIS; GABRILOVICH, 2014). Rzeszotarski e Kittur (2011) analisam em que medida a acurácia de um trabalhador se relaciona com métricas comportamentais que indicam sinais de engajamento de curto prazo, tais como: número de cliques, total de tempo trabalhando na tarefa e total de movimentos do *mouse*. Nas tarefas avaliadas, há uma correlação entre as acurácias preditas usando esses tipos de métricas e as acurácias reais dos trabalhadores. O que se obteve na pesquisa descrita neste comento é que métricas de engajamento de curto prazo se relacionam com a credibilidade. Essa relação varia muito com o projeto e com o trabalhador e a dificuldade das tarefas desempenha um papel muito importante na análise dessa relação.

Também nesse contexto, Ipeirotis e Gabrilovich (2014) implementam e avaliam o sistema Quizz. A ênfase da avaliação está na escalabilidade do sistema a até um bilhão de participantes. Entretanto, na avaliação realizada com 4.091 participantes em setembro de 2013, eles reportaram a ocorrência de um fenômeno definido como auto seleção (*self-selection*). O que caracteriza esse fenômeno é que trabalhadores que apresentam respostas com baixa qualidade tendem a executar poucas tarefas e, em seguida, deixarem o sistema. Esse resultado pode ser visto de diversas perspectivas. Por exemplo, cabe questionar se os trabalhadores que apresentam baixa acurácia e tendem deixar o sistema são aqueles que sentem mais dificuldade para executar as tarefas ou são aqueles que têm pouco interesse em executar as tarefas adequadamente.

8.4 Replicação

Replicação de tarefas não é algo que tem recebido atenção em computação por humanos. O que se pratica é o uso de replicação fixa com o objetivo de obter redundância para posterior

identificação e remoção de respostas discrepantes por meio da agregação. Nessa abordagem, o nível de replicação é quase sempre superestimado. Há casos, por exemplo, em que uma mesma tarefa foi replicada para 30 trabalhadores diferentes (LINTOTT et al., 2008).

Como discutido nas últimas seções, diversos estudos destacam a necessidade e o desafio de atrair e engajar trabalhadores. Entretanto, os resultados obtidos neste trabalho mostram que muitas vezes o sistema não faz um uso adequado do poder computacional provido por aqueles que se engajam. Muito do poder computacional provido por eles é gasto com redundância muitas vezes desnecessária. Obteve-se que um algoritmo de replicação adaptativo pode economizar réplicas em relação à replicação fixa sem comprometer a acurácia.

Outra vantagem do algoritmo proposto em relação ao estado da arte é facilitar a identificação das tarefas consideradas "sem conclusão". Trata-se de tarefas nas quais a replicação foi interrompida por que o número máximo de réplicas foi atingido, mas sem que um grupo de respostas tenha atingido o limiar de credibilidade requerido pelo usuário. Em computação por humanos, esse tipo de tarefa pode ter grande importância para o usuário. Os dados de entrada em uma tarefa sem conclusão podem revelar algo fora do padrão que o usuário pode ter interesse em investigar com maior cuidado ou submeter à análise de trabalhadores mais especializados. Dessa forma, a estratégia proposta também facilita a análise dos dados obtidos do sistema.

Replicação tem sido utilizada em outros tipos de sistemas distribuídos cujas semelhanças podem ser discutidas (SARMENTA, 2002; CIRNE et al., 2007). Em sistemas de computação voluntária implementados usando o sistema BOINC[1], replicação é utilizada como forma de identificar "máquinas sabotadoras" e isolar as respostas providas por elas (SARMENTA, 2002). Isso se justifica dado que, neste tipo de sistema, se uma máquina gerar uma resposta divergente das demais, acima de uma margem de erro, esse fato é um indicador de que ela está em um estado errôneo ou que ela é uma máquina que gera resultados incorretos de forma intencional. Seres humanos, no entanto, podem prover uma resposta incorreta em razão de diversos fatores das instruções da tarefa, dos dados de entrada da tarefa e até mesmo fatores do ambiente no qual ele se encontra (KOCHHAR; MAZZOCCHI; PARITOSH, 2010; EICKHOFF; VRIES, 2011). Além disso, o conceito de corretude não se aplica a todos os tipos de tarefas de

[1] Acrônimo da expressão em inglês *Berkeley Open Infrastructure for Network Computing*. Trata-se de um sistema de *middleware* para sistemas de computação voluntária (ANDERSON, 2004).

computação por humanos. Dessa forma, enquanto a principal questão tratada por Sarmenta era analisar a quantidade de tempo que o sistema gasta para detectar e eliminar uma máquina sabotadora, neste trabalho a principal questão é como obter uma resposta crível com o mínimo de réplicas e levando em conta o fato de que os trabalhadores estão sujeitos a ignorância, esquecimento e deslize que podem ter diversas causas. Isso é realizada considerando a dificuldade das tarefas e uma diversidade de métricas de credibilidade que levam em conta diferentes fatores das tarefas e dos trabalhadores.

Também é importante destacar que os sistemas de computação voluntária baseados em máquinas serviram de inspiração para diversos sistemas de computação por humanos. Esse é o caso, por exemplo, dos sistemas Bossa[2] e PyBossa[3], nos quais uma das principais diferenças em relação a sistemas de computação voluntária é levar em conta fatores humanos dos trabalhadores nas estratégias projetadas para se obter melhor desempenho. Até onde se sabe, nenhum desses sistemas implementa estratégias de replicação de tarefas de forma adaptativa que permitam otimizar o número de réplicas levando em conta, por exemplo, informações da credibilidade dos trabalhadores e estimativas de dificuldade das tarefas.

Um tipo de replicação que pode ser destacado é o empregado em ambientes de computação oportunista, como as grades computacionais oportunistas (SILVA; CIRNE; BRASILEIRO, 2003; CIRNE et al., 2007). Nessas grades, uma mesma tarefa é replicada em diversas máquinas com o objetivo de que alguma delas termine a execução mais rápido. Apenas a resposta da primeira réplica que terminar a execução é utilizada e as demais réplicas são canceladas (SILVA; CIRNE; BRASILEIRO, 2003). Assim, já se pressupõe que algumas réplicas serão canceladas e haverá desperdício de poder computacional. Com essa estratégia, esse tipo de sistema obtém uma redução no tempo de resposta das aplicações. A replicação proposta na pesquisa descrita neste documento tem o objetivo de tolerar falhas, mas com o mínimo de desperdício de poder cognitivo. Buscou-se controlar a geração das réplicas de modo que nenhuma réplica seja gerada sem necessidade. Essa preocupação com o número de réplicas geradas, que não é crítico em grades computacionais, é crucial em computação por huma-

[2]Acrônimo da expressão em inglês *BOINC Open System for Skill Aggregation* (https://boinc.berkeley.edu/trac/wiki/BossaIntro, último acesso em 01 de outubro de 2015). Trata-se de um sistema de *middleware* para sistemas de computação por humanos (ANDERSON, 2008). Esse sistema é inspirado no sistema de computação voluntária BOINC.

[3]O sistema PyBossa (www.pybossa.org, último acesso em 01 de outubro de 2015) surgiu como uma versão do sistema Bossa implementada na linguagem de programação Python. Entretanto, atualmente é um sistema independente e reúne diversas funcionalidades que não existem no Bossa (GONZÁLEZ et al., 2015).

nos, pois, como discutido ao longo deste documento, nesses sistemas o engajamento dos trabalhadores tende a ser baixo. Em razão disso, a utilidade de cada resposta provida por um trabalhador deve ser maximizada.

Uma semelhança que pode ser apontada nesse contexto se refere ao requisito de tempo. Em ambientes em que a computação é realizada de forma oportunista, admite-se um desperdício de poder computacional para que se tenha um ganho de paralelismo. Isso também ocorre na estratégia de replicação descrita neste documento. O usuário pode alterar o parâmetro de urgência de modo a aumentar o paralelismo na execução das tarefas. Mostrou-se, entretanto, que aumentar demais a urgência pode acarretar, além do aumento do desperdício de poder computacional, outros custos como o aumento na proporção de tarefas para as quais uma resposta crível não pôde ser obtida.

Capítulo 9

Conclusões

O objeto de estudo da pesquisa reportada neste documento foi a oferta de poder cognitivo em sistemas de computação por humanos. Propôs-se como objetivo geral investigar a tese de que um entendimento maior do engajamento e da credibilidade dos seres humanos que atuam no sistema permite caracterizar a oferta de poder cognitivo e implementar estratégias que otimizem o uso que o sistema faz do poder cognitivo disponível. Investigou-se o papel da replicação de tarefas baseada em credibilidade para se atingir tal otimização de desempenho. Neste capítulo, os principais resultados e contribuições são ressaltados (Seção 9.1) e diversas perspectivas de futuras pesquisas são apresentadas (Seção 9.2).

9.1 Resultados e Contribuições

Pode-se destacar três principais contribuições deste trabalho. A primeira contribuição foi a articulação de um arcabouço conceitual sobre o qual computação por humanos pode ser analisado em uma perspectiva de sistemas distribuídos (PONCIANO et al., 2014). Esse arcabouço teve como propósito evidenciar os principais agentes no ecossistema de computação por humanos, que foram identificados como sendo três principais: os usuários, os trabalhadores e o sistema que intermedia a interação entre eles. Tal visão do ecossistema trouxe à luz os principais componentes em um sistema de computação por humanos que são considerados na análise e otimização de desempenho. São três esses componentes: os requisitos de qualidade de serviço dos usuários, os aspectos humanos dos trabalhadores e as estratégias de projeto e gerência de aplicações implementadas pelo sistema. Até onde se sabe, esse

arcabouço é um dos primeiros que tenta prover uma perspectiva de sistema distribuído no contexto de computação por humanos. Além de servir a diversos propósitos, ele mostrou-se especialmente valioso para orientar o estudo da oferta e otimização de poder cognitivo no sistema.

A segunda contribuição deste trabalho foi a pesquisa de formas de caracterizar a oferta de poder cognitivo em sistemas de computação por humanos (PONCIANO et al., 2014b; PONCIANO; BRASILEIRO, 2014). Nessa pesquisa, utilizou-se os arcabouços teórico-conceituais de engajamento e credibilidade que têm uso multidisciplinar e que consideram tanto aspectos humanos quanto aspectos gerais de computação. O arcabouço de engajamento de seres humanos informou a proposta de quatro métricas para medir a duração e o grau de engajamento dos trabalhadores no sistema: duração relativa da atividade, variação na periodicidade, taxa de atividade, tempo dedicado diariamente. O arcabouço de credibilidade informou a proposta de 4 métricas de credibilidade: concordância simples, concordância reputada, concordância experimentada e concordância ponderada.

Finalmente, a terceira contribuição consistiu na proposta de um algoritmo de replicação adaptativo. Esse algoritmo implementa uma replicação ativa em que o nível de replicação é definido de forma adaptativa (PONCIANO; BRASILEIRO; GADELHA, 2013; PONCIANO et al., 2014a). O principal objetivo do algoritmo é melhorar o uso que o sistema faz do poder cognitivo provido pelos trabalhadores. A principal ideia do algoritmo é identificar, para cada tarefa, quando uma resposta suficientemente crível é obtida para que mais réplicas da tarefa não precisem ser geradas. Isso é feito pela otimização do nível de replicação levando em conta características de credibilidade dos trabalhadores, de dificuldade das tarefas e requisitos de qualidade de serviço do usuário, como grau de credibilidade requerida nas respostas e urgência.

A avaliação das métricas de engajamento e de credibilidade se deu por meio de estudos de caso de caracterização do engajamento e da credibilidade dos trabalhadores em seis projetos de computação por humanos bastante distintos: Galaxy Zoo, The Milky Way Project, Cell Spotting, Sun4All, Análise de Sentimentos e Julgamento de Fatos. Tal caracterização envolveu a análise de semelhanças e diferenças entre os trabalhadores e de relações entre métricas por meio de análises de classificação, agrupamento, correlação e regressões. A avaliação do algoritmo de replicação, por sua vez, consistiu em simulações que permitiram medir a ca-

pacidade do algoritmo proposto de otimizar o número de réplicas utilizadas e a acurácia das respostas considerando a variação de diversos parâmetros de interesse como nível de credibilidade requerida na resposta final, urgência de obter uma resposta e métrica de credibilidade. A análise de configurações mais adequadas se deu por otimização multiobjetivo no conjunto de configurações avaliadas.

Os resultados obtidos revelam diversas características da oferta de poder cognitivo em computação por humanos. Características tais que até então permaneciam desconhecidas. Considerando as métricas de engajamento, observou-se que existem duas grandes classes de trabalhadores: transientes e regulares. Os trabalhadores transientes constituem uma maioria pouca engajada. Os trabalhadores regulares são minoritários, mas eles apresentam maior contribuição em termos tanto da proporção de tarefas executadas quanto do total de tempo de computação disponibilizado ao sistema. Trabalhadores nessa classe exibem 5 perfis de engajamento cognitivo, que podem ser rotulados como: empenhados, espasmódicos, persistentes, duradouros e moderados. Cada perfil de engajamento cognitivo representa um padrão de provimento de poder cognitivo que se destaca de forma positiva ou negativa em relação aos demais perfis em alguma métrica de engajamento.

O estudo da credibilidade dos trabalhadores permitiu analisar diferentes formas de se estimar o quanto se pode acreditar em cada resposta obtida de trabalhadores em sistemas de computação por humanos. Observou-se que a credibilidade deles pode ser medida usando diferentes métricas baseadas no nível de concordância entre eles. Naturalmente, o valor da credibilidade tende a variar com a métrica utilizada, dependendo se ela é mais conservadora (como a concordância reputada) ou menos conservadora (como a concordância simples). A ordem de credibilidade dos trabalhadores também muda dependendo da métrica de credibilidade utilizada. Pode-se destacar também que a credibilidade de cada trabalhador varia com a dificuldade da tarefa que ele executa e que os trabalhadores tendem a apresentar maior variação de credibilidade entre eles em tarefas de dificuldade moderada e difícil.

Por fim, o estudo da replicação de tarefas mostrou que é possível usar melhor o poder cognitivo provido pelos trabalhadores. A replicação adaptativa permite que isso seja feito atendendo os requisitos de credibilidade requerida e de urgência do usuário. Ao considerar esses requisitos, estratégias de otimização mostram-se úteis na definição dos valores a serem utilizados nos parâmetros do algoritmo. No geral, o algoritmo de replicação proposto

apresenta desempenho superior à replicação com nível de replicação fixo e que usa voto majoritário para eleger a resposta final para cada tarefa. O desempenho do algoritmo pode se aproximar do desempenho de um oráculo que sabe se um trabalhador proverá uma resposta correta ou incorreta.

9.2 Trabalhos Futuros

A pesquisa reportada neste documento não é exaustiva em estudar arcabouços teóricos de engajamento e credibilidade nem em extrair métricas quantitativas de tais arcabouços. Muitas pesquisas ainda podem ser conduzidas nessa direção. Trabalhos futuros podem considerar a análise de outros arcabouços que se mostrem úteis na análise da oferta de poder cognitivo em computação por humanos. A extração de novas métricas dos arcabouços utilizados neste documento também pode ser considerada. Nesses casos, uma análise do ganho de informação em relação ao estudo conduzido neste trabalho é importante. Ou seja, é importante colocar em perspectiva que tipo de conhecimento novo sobre a oferta de poder cognitivo se pode obter ao utilizar outros arcabouços e métricas.

Considerando os resultados reportados neste documento e as pesquisas que têm sido realizadas por trabalhos relacionados, mostra-se relevante a condução de mais estudos qualitativos que investiguem relações entre os motivos do engajamento e da credibilidade reportados pelos trabalhadores em questionários e entrevistas e o engajamento e a credibilidade tal como caracterizados em dados de atuação deles no sistema. Em um primeiro momento, esse tipo de estudo se mostra importante para explicar os fatores humanos que determinam a credibilidade e o engajamento que se manifestam no sistema. Por exemplo, um fator de grande importância é entender em que situações a dificuldade da tarefa é percebida pelo trabalhador como um desafio e, portanto, aumenta o engajamento, e em que situações ela é percebida como um sinal de que ele está fazendo algo errado e, portanto, reduz seu engajamento. Em um segundo momento, esse tipo de estudo pode orientar o desenvolvimento de uma nova geração de estratégias que visem melhorar a experiência dos trabalhadores no sistema. Tais estratégias devem não apenas detectar e se adaptar ao comportamento detectado no histórico de execução de tarefas, mas incorporar as métricas na dinâmica do sistema de modo a colocar em prática ações que determinem o comportamento futuro dos trabalhadores.

Como exemplo de incorporação das métricas na dinâmica do sistema, pode-se considerar estratégias personalizadas de engajamento. Ao mostrar que os trabalhadores em sistemas de computação por humanos se comportam de forma muito diferente entre si, este estudo motiva o desenvolvimento de um componente de gerenciamento do engajamento dos trabalhadores. Tal componente pode monitorar o comportamento de cada trabalhador, e, quando necessário, ativar automaticamente uma estratégia de engajamento adequada. Trabalhadores que se comportam de forma diferente devem ser sujeitos a estratégias de engajamento diferentes. As estratégias podem se concentrar em promover a redução do engajamento ou o aumento do engajamento. Estratégias podem se concentrar em promover o aumento do engajamento dos trabalhadores quando eles apresentam um nível de engajamento abaixo do nível típico no projeto, considerando as quatro métricas propostas. Estratégias de redução do engajamento podem ser importantes quando alguns trabalhadores começam a comprometer muito do seu tempo executando tarefas no projeto, o que pode levar a perdas em suas interações sociais ou até mesmo esgotamento (MASLACH; JACKSON, 1981). Nessa mesma linha de estratégias personalizadas, também se mostra relevante projetar estratégias de composição de aplicações, atribuição de tarefas, incentivos e retribuições.

Os resultados obtidos neste trabalho revelaram relações entre engajamento, dificuldade e credibilidade. Em um contexto semelhante, trabalhos futuros também devem investigar em maior profundidade a evolução do comportamento dos trabalhadores nos sistemas, inspirado em estudos conduzidos em outros tipos de sistemas (WEN; ROSE, 2012; FURTADO et al., 2013; VAROL et al., 2014). Nesse contexto há diversas questões de interesse, por exemplo: (i) que fluxos de comportamento podem levar um trabalhador a se tornar engajado e crível? (ii) que fluxos podem levá-lo a um desengajamento prematuro e características de baixa credibilidade? e (ii) os trabalhadores que se tornaram persistentes apresentaram um fluxo de comportamento semelhante? Neste contexto, também se pode investigar a transferência de conhecimento ao longo de diferentes graus de dificuldade de tarefas. Por exemplo, trabalhadores que se mostraram críveis em tarefas do grau de dificuldade n tendem a apresentar que nível de credibilidade em tarefas do grau de dificuldade $n + 0,1$? Esse tipo de informação pode permitir adaptações no algoritmo de replicação que melhorem ainda mais o seu desempenho.

Naturalmente as análises conduzidas neste trabalho podem ser ampliadas para contextos

ainda mais complexos que tendem a ganhar importância com o aumento do uso de computação por humanos. Três cenários pouco comuns nos dias atuais, mas com tendência de crescimento são: (i) existência de múltiplos sistemas de computação por humanos que podem ser escolhidos pelos usuários para executar suas tarefas, (ii) existência de sistemas que agregam múltiplos projetos, e (iii) sistemas que combinem computação por humanos e computação por máquinas.

Com o crescimento da quantidade de sistemas de computação por humanos, os usuários poderão escolher em que sistema executar as suas aplicações. Neste contexto, mostra-se de grande importância medir a credibilidade do sistema em si, não apenas a credibilidade dos trabalhadores que atuam no sistema como investigado na pesquisa a que se refere este documento. Por exemplo, em termos de credibilidade, como comparar os sistemas de pensamento voluntário Zooniverse e Crowdcrafting ou os mercados Amazon Mechanical Turk e Crowflower? A medição da credibilidade e do engajamento dos trabalhadores que atuam em cada sistema pode se mostrar relevante, mas não suficiente. Por exemplo, para medir a credibilidade de um sistema, pode-se mostrar importante considerar as características dos diversos tipos de estratégias implementadas por ele (PONCIANO et al., 2014).

Quanto à ampliação das análises no contexto de sistemas com múltiplos projetos, trata-se de uma questão inerente de sistemas de pensamento voluntário. Uma das promessas dos sistemas com múltiplos projetos é facilitar o recrutamento de trabalhadores que tendem a ser engajados e críveis (FORTSON et al., 2012). O argumento é que reunir em um só lugar projetos que apresentam semelhanças e que possam somar esforços uns aos outros permite, entre outras vantagens, reduzir o custo de recrutar e engajar trabalhadores. No entanto, em que medida os projetos realmente se beneficiam desses sistemas ainda é algo que permanece desconhecido. Existem duas questões chaves neste contexto: (i) em que medida trabalhadores recrutados por um projeto migram para outros projetos e (ii) em que medida os trabalhadores que migram realmente se engajam nos outros projetos. Como forma de fomentar esse estudo, algumas análises preliminares dessas questões foram conduzidas no sistema Crowdcrafting e são reportadas no Apêndice C.

Há uma tendência de que surja uma nova geração de sistemas de computação que combinem computação por humanos e computação por máquinas (JENNINGS et al., 2014). Nesses sistemas, seres humanos e máquinas podem ser avaliadores uns dos outros ou complemen-

tares executando tarefas diferentes. Também nesses sistemas, o estudo de engajamento e credibilidade se mostra necessário. Entender a oferta de poder cognitivo por meio do engajamento pode se mostrar fundamental para evitar períodos de inércia do sistema, por exemplo, máquinas aguardando um dado que deve ser gerado por um ser humano. Entender a credibilidade nesse contexto pode se mostrar importante para prevenir, identificar e tratar eventuais falhas a que máquinas e seres humanos estão sujeitos.

Por fim, o estudo da oferta de poder cognitivo em sistemas de computação por humanos reportado neste documento mostrou-se uma fundação sólida. Sobre essa fundação, diversos outros estudos podem ser conduzidos. Espera-se que este trabalho inspire novas pesquisas em computação por humanos, em especial em credibilidade, engajamento e replicação de tarefas.

Bibliografia

AHN, L. *Human computation*. Tese (Doutorado) — Carnegie Mellon University, 2005. UMU Order Number: AAI3205378.

AHN, L. V. et al. reCAPTCHA: Human-based character recognition via Web security measures. *Science*, 2008. American Association for the Advancement of Science, USA, v. 321, n. 5895, p. 1465–1468, 2008. ISSN 1095-9203.

AHN, L. von; DABBISH, L. Labeling images with a computer game. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2004. p. 319–326. ISBN 1-58113-702-8.

AMBATI, V.; VOGEL, S.; CARBONELL, J. G. Towards task recommendation in micro-task markets. In: *AAAI Workshop on Human Computation*. Palo Alto, CA, USA: AAAI, 2011. p. 80–83. Disponível em: <http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/4005>. Acesso em: 1 de outubro de 2015.

AMIR, O. et al. On the verification complexity of group decision-making tasks. In: *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*. Palo Alto, CA, USA: AAAI, 2013. p. 2–8. ISBN 978-1-57735-607-3. Disponível em: <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7488>. Acesso em: 1 de outubro de 2015.

ANDERBERG, M. *Cluster analysis for applications*. Waltham, Massachusetts,United States: Academic Press, 1973. ISBN 978-0-12057-650-0.

ANDERSON, D. Boinc: a system for public-resource computing and storage. In: *Fifth IEEE/ACM International Workshop on Grid Computing*. Washington, DC, USA: IEEE, 2004. p. 4–10. ISSN 1550-5510.

ANDERSON, D. P. *Bossa: Middleware for Volunteer Thinking*. September 2008. Disponível em: <http://boinc.berkeley.edu/talks/bossa_intro.pdf>. Acesso em: 1 de outubro de 2015.

ARAÚJO, R. 99designs: An analysis of creative competition in crowdsourced design. In: *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*. Palo Alto, CA, USA: AAAI, 2013. p. 17–24. Disponível em: <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7519/7399>. Acesso em: 1 de outubro de 2015.

ARCANJO, J. et al. Evaluating volunteers' contributions in a citizen science project. In: *Proceedings of the 10th IEEE International Conference on e-Science (e-Science)*. Washington, DC, USA: IEEE, 2014. v. 1, p. 21–28.

ARCHAK, N. Money, glory and cheap talk: Analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on topcoder.com. In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: ACM, 2010. p. 21–30. ISBN 978-1-60558-799-8.

ARGUELLO, J. et al. Talk to me: Foundations for successful individual-group interactions in online communities. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2006. p. 959–968. ISBN 1-59593-372-7.

ARLITT, M. Characterizing Web user sessions. *SIGMETRICS Perform. Eval. Rev.*, 2000. ACM, New York, NY, USA, v. 28, n. 2, p. 50–63, set. 2000. ISSN 0163-5999.

AROYO, L.; WELTY, C. The three sides of CrowdTruth. *Human Computation*, 2014. v. 1, n. 1, p. 31–44, 2014. ISSN 2330-8001.

ASLAM, J.; PAVLU, V. Query hardness estimation using jensen-shannon divergence among multiple scoring functions. In: AMATI, G.; CARPINETO, C.; ROMANO, G. (Ed.). *Advances in Information Retrieval*. Berlin, Germany: Springer, 2007, (Lecture Notes in Computer Science, v. 4425). p. 198–209. ISBN 978-3-540-71494-1.

ATTFIELD, S. et al. Towards a science of user engagement. In: *WSDM Workshop on User Modelling for Web Applications*. New York, NY, USA: ACM, 2011. p. 1–8.

BAKKER, A. B.; DEMEROUTI, E. Towards a model of work engagement. *Career development international*, 2008. Emerald Group Publishing Limited, v. 13, n. 3, p. 209–223, 2008. ISSN 1362-0436.

BANDURA, A. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 1977. American Psychological Association, v. 84, n. 2, p. 191, 1977.

BAROWY, D. W. et al. Automan: A platform for integrating human-based and digital computation. *SIGPLAN Not.*, 2012. ACM, New York, NY, USA, v. 47, n. 10, p. 639–654, out. 2012. ISSN 0362-1340.

BEIRNE, C.; LAMBIN, X. Understanding the determinants of volunteer retention through capture-recapture analysis: Answering social science questions using a wildlife ecology toolkit. *Conservation Letters*, 2013. v. 6, n. 6, p. 391–401, 2013. ISSN 1755-263X.

BERNSTEIN, A.; KLEIN, M.; MALONE, T. W. Programming the global brain. *Commun. ACM*, 2012. ACM, New York, NY, USA, v. 55, n. 5, p. 41–43, maio 2012. ISSN 0001-0782.

BERNSTEIN, M. S. et al. Soylent: A word processor with a crowd inside. In: *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 2010. p. 313–322. ISBN 978-1-4503-0271-5.

BIRD, T. J. et al. Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, 2014. Elsevier Science Publishers B. V., Amsterdam, Netherlands, v. 173, p. 144 – 154, 2014. ISSN 0006-3207.

BOZZON, A. et al. Reactive crowdsourcing. In: *Proceedings of the 22nd International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. p. 153–164. ISBN 978-1-4503-2035-1.

BRAGG, J.; MAUSAM; WELD, D. S. Crowdsourcing multi-label classification for taxonomy creation. In: *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*. Palo Alto, CA, USA: AAAI, 2013. p. 25–33. ISBN 978-1-57735-607-3. Disponível em: <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7560>. Acesso em: 1 de outubro de 2015.

BRANSON, S. et al. Visual recognition with humans in the loop. In: *Proceedings of the 11th European Conference on Computer Vision: Part IV*. Berlin, Heidelberg: Springer-Verlag, 2010. p. 438–451. ISBN 3-642-15560-X, 978-3-642-15560-4.

CALLISTER, R.; SUWARNO, N. O.; SEALS, D. R. Sympathetic activity is influenced by task difficulty and stress perception during mental challenge in humans. *The Journal of Physiology*, 1992. v. 454, n. 1, p. 373–387, 1992. ISSN 1469-7793.

CARDOSO, J. et al. Modeling quality of service for workflows and Web service processes. *J Web Semant*, 2002. Elsevier Science Publishers B. V., Amsterdam, Netherlands, v. 1, p. 281–308, 2002. ISSN 1570-8268.

CERUZZI, P. E. When computers were human. *Annals of the History of Computing*, 1991. IEEE, Washington, DC, USA, v. 13, n. 3, p. 237–244, July 1991. ISSN 0164-1239.

CHANDLER, D.; HORTON, J. J. Labor allocation in paid crowdsourcing: Experimental evidence on positioning, nudges and prices. In: *AAAI Human Computation workshop*. Palo Alto, CA, USA: AAAI, 2011. p. 14–19. Disponível em: <https://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3983>. Acesso em: 1 de outubro de 2015.

CHILTON, L. B. et al. Task search in a human computation market. In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*. New York, NY, USA: ACM, 2010. p. 1–9. ISBN 978-1-4503-0222-7.

CHILTON, L. B. et al. Cascade: Crowdsourcing taxonomy creation. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2013. p. 1999–2008. ISBN 978-1-4503-1899-0.

CHRISTENSEN, L. *Experimental Methodology*. New Yourk, USA: Pearson/Allyn & Bacon, 2007. (Pearson International Edition). ISBN 9780205484737.

CIRNE, W. et al. On the efficacy, efficiency and emergent behavior of task replication in large distributed systems. *Parallel Comput.*, 2007. Elsevier Science Publishers B. V., Amsterdam, Netherlands, v. 33, n. 3, p. 213–234, abr. 2007. ISSN 0167-8191.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960. Sage, v. 20, n. 1, p. 37–46, abr. 1960.

COLEMAN, J. *Foundations of social theory*. Cambridge, Massachusetts, United States: Harvard, 1990. ISBN 0-674-31225-2.

CONLISK, J. Why bounded rationality? *Journal of economic literature*, 1996. American Economic Association, v. 34, n. 2, p. 669–700, 1996. Disponível em: <http://www.jstor.org/stable/2729218>. Acesso em: 1 de outubro de 2015.

COULOURIS, G.; DOLLIMORE, J.; KINDBERG, T. *Distributed Systems (4rd Ed.): Concepts and Design*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0-321-26354-5.

CROUSER, R.; CHANG, R. An affordance-based framework for human computation and human-computer collaboration. *IEEE Transactions on Visualization and Computer Graphics*, 2012. IEEE, Washington, DC, USA, v. 18, n. 12, p. 2859–2868, Dec 2012. ISSN 1077-2626.

DALVI, N. et al. Aggregating crowdsourced binary ratings. In: *Proceedings of the 22nd International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. p. 285–294. ISBN 978-1-4503-2035-1.

DECI, E. L.; RYAN, R. M. The "what"and "why"of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 2000. Taylor & Francis, v. 11, n. 4, p. 227–268, 2000.

DIFALLAH, D. E.; DEMARTINI, G.; CUDRÉ-MAUROUX, P. Pick-a-crowd: Tell me what you like, and I'll tell you what to do. In: *Proceedings of the 22nd International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. p. 367–374. ISBN 978-1-4503-2035-1.

DOLAN, R. J. Emotion, cognition, and behavior. *Science*, 2002. American Association for the Advancement of Science, USA, v. 298, n. 5596, p. 1191–1194, 2002.

EICKHOFF, C.; VRIES, A. de. How crowdsourcable is your task? In: LEASE, M.; CARVALHO, V.; YILMAZ, E. (Ed.). *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*. New York, NY, USA: ACM, 2011. p. 11–14.

EVELEIGH, A. et al. Designing for dabblers and deterring drop-outs in citizen science. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2014. p. 2985–2994. ISBN 978-1-4503-2473-1.

FAGIN, R.; KUMAR, R.; SIVAKUMAR, D. Comparing top k lists. In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2003. p. 28–36. ISBN 0-89871-538-5.

FERREIRA, A.; ANJOS, M. dos. *Dicionário Aurélio básico da língua portuguesa*. Rio de Janeiro, Brasil: Editora Nova Fronteira, 1988. ISBN 9-788-52090-826-6.

FISCHER, G. User modeling in human-computer interaction. *User Modeling and User-Adapted Interaction*, 2001. Kluwer Academic Publishers, Hingham, MA, USA, v. 11, n. 1-2, p. 65–86, mar. 2001. ISSN 1573-1391.

FLEISS, J. L.; LEVIN, B.; PAIK, M. C. The measurement of interrater agreement. In: *Statistical methods for rates and proportions*. New York, NY, USA: John Wiley and Sons, 1981. v. 2, p. 212–236.

FOGG, B. J. Prominence-interpretation theory: Explaining how people assess credibility online. In: *Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2003. p. 722–723. ISBN 1-58113-637-4.

FOGG, B. J.; TSENG, H. The elements of computer credibility. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 1999. p. 80–87. ISBN 0-201-48559-1.

FORGY, E. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 1965. International Biometric Society, v. 21, n. 3, p. 768–780, 1965.

FORTSON, L. et al. Galaxy Zoo: Morphological classification and citizen science. In: *Advances in Machine Learning and Data Mining for Astronomy*. Boca Raton, Florida, USA: CRC Press, 2012. p. 213–236.

FURTADO, A. et al. Contributor profiles, their dynamics, and their importance in five Q&A sites. In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, 2013. p. 1237–1252. ISBN 978-1-4503-1331-5.

GEIGER, R. S.; HALFAKER, A. Using edit sessions to measure participation in Wikipedia. In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, 2013. p. 861–870. ISBN 978-1-4503-1331-5.

GONZÁLEZ, D. et al. *pybossa: v0.2.2*. May 2015. Disponível em: <http://dx.doi.org/10.5281/zenodo.17516>. Acesso em: 1 de outubro de 2015.

GONZÁLEZ-ROMÁ, V. et al. Burnout and work engagement: Independent factors or opposite poles? *Journal of Vocational Behavior*, 2006. Elsevier Science Publishers B. V., Amsterdam, Netherlands, v. 68, n. 1, p. 165–174, 2006.

GREITZER, F. L. Toward the development of cognitive task difficulty metrics to support intelligence analysis research. In: *Proceedings of the Fourth IEEE International Conference on Cognitive Informatics*. Washington, DC, USA: IEEE Computer Society, 2005. p. 315–320. ISBN 0-7803-9136-5.

GRIER, D. A. *When computers were human*. Princeton, New Jersey, USA: Princeton University Press, 2007. ISBN 978-1-40-084936-9.

GRUDIN, J.; POLTROCK, S. Taxonomy and theory in computer supported cooperative work. In: KOZLOWSKI, S. W. J. (Ed.). *Handbook of organizational psychology*. New York: Oxford University Press, 2012. p. 1323–1348.

HOVY, D. et al. Learning whom to trust with MACE. In: *Conference of the North American Chapter of the Association of Computational Linguistics*. The Association for Computational Linguistics, 2013. p. 1120–1130. Disponível em: <http://aclweb.org/anthology/N/N13/N13-1132.pdf>. Acesso em: 1 de outubro de 2015.

HOWE, J. The rise of crowdsourcing. *Wired magazine*, 2006. v. 14, n. 6, p. 1–4, 2006.

HU, C.; BEDERSON, B. B.; RESNIK, P. Translation by iterative collaboration between monolingual users. In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*. New York, NY, USA: ACM, 2010. p. 54–55. ISBN 978-1-4503-0222-7.

IPEIROTIS, P. G. Analyzing the Amazon Mechanical Turk marketplace. *XRDS*, 2010. ACM, New York, NY, USA, v. 17, n. 2, p. 16–21, dez. 2010. ISSN 1528-4972.

IPEIROTIS, P. G.; GABRILOVICH, E. Quizz: Targeted crowdsourcing with a billion (potential) users. In: *Proceedings of the 23rd International Conference on World Wide Web*. New York, NY, USA: ACM, 2014. p. 143–154. ISBN 978-1-4503-2744-2.

IPEIROTIS, P. G.; PROVOST, F.; WANG, J. Quality management on Amazon Mechanical Turk. In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*. New York, NY, USA: ACM, 2010. p. 64–67. ISBN 978-1-4503-0222-7.

IRANI, L. C.; SILBERMAN, M. S. Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2013. p. 611–620. ISBN 978-1-4503-1899-0.

JAIN, R. *The art of computer systems performance analysis*. Hoboken, New Jersey, US: John Wiley & Sons, 2008. ISBN 0-471-50336-3.

JALOTE, P. *Fault tolerance in distributed systems*. New Jersey, USA: Prentice Hall, 1994. ISBN 978-0-13-301367-2.

JENNINGS, N. R. et al. Human-agent collectives. *Communications of the ACM*, 2014. ACM, New York, NY, USA, v. 57, n. 12, p. 80–88, nov. 2014. ISSN 0001-0782.

JUFFINGER, A.; GRANITZER, M.; LEX, E. Blog credibility ranking by exploiting verified content. In: *Proceedings of the 3rd Workshop on Information Credibility on the Web*. New York, NY, USA: ACM, 2009. p. 51–58. ISBN 978-1-60558-488-1.

KAFRI, R.; LEVY, M.; PILPEL, Y. The regulatory utilization of genetic redundancy through responsive backup circuits. *Proceedings of the National Academy of Sciences*, 2006. v. 103, n. 31, p. 11653–11658, 2006.

KAPTEYN, A.; WANSBEEK, T.; BUYZE, J. The dynamics of preference formation. *Economics Letters*, 1978. v. 1, n. 1, p. 93 – 98, 1978. ISSN 0165-1765.

KAZAI, G.; KAMPS, J.; MILIC-FRAYLING, N. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 2013. Springer Netherlands, v. 16, n. 2, p. 138–178, 2013. ISSN 1386-4564.

KEARNS, M. Experiments in social computation. *Commun. ACM*, 2012. ACM, New York, NY, USA, v. 55, n. 10, p. 56–67, out. 2012. ISSN 0001-0782.

KHANNA, S. et al. Evaluating and improving the usability of Mechanical Turk for low-income workers in India. In: *Proceedings of the First ACM Symposium on Computing for Development*. New York, NY, USA: ACM, 2010. p. 12:1–12:10. ISBN 978-1-4503-0473-3.

KITTUR, A. et al. The future of crowd work. In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, 2013. p. 1301–1318. ISBN 978-1-4503-1331-5.

KOCHHAR, S.; MAZZOCCHI, S.; PARITOSH, P. The anatomy of a large-scale human computation engine. In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*. New York, NY, USA: ACM, 2010. p. 10–17. ISBN 978-1-4503-0222-7.

KRIPPENDORFF, K. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 1970. Sage Publications, v. 30, n. 1, p. 61–70, 1970.

KULKARNI, A.; CAN, M.; HARTMANN, B. Collaboratively crowdsourcing workflows with Turkomatic. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, 2012. p. 1003–1012. ISBN 978-1-4503-1086-4.

LAW, E. Defining (human) computation. In: *ACM CHI 2001 workshop on Crowdsourcing and Human Computation*. New York, NY, USA: ACM, 2011. p. 1–4. Disponível em: <http://crowdresearch.org/chi2011-workshop/papers/law.pdf>. Acesso em: 1 de outubro de 2015.

LAW, E.; AHN, L. von. *Human Computation*. California, USA: Morgan & Claypool Publishers, 2011. (Synthesis Lectures on Artificial Intelligence and Machine Learning). ISBN 978-1-60845-517-1.

LEHMANN, J. et al. Models of user engagement. In: *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*. Berlin, Heidelberg: Springer-Verlag, 2012. p. 164–175. ISBN 978-3-642-31453-7.

LIGHT, J. S. When computers were women. *Technology and Culture*, 1999. The Johns Hopkins University Press, Baltimore, Maryland, USA, v. 40, n. 3, p. 455–483, 1999.

LIN, C. H.; MAUSAM; WELD, D. S. Dynamically switching between synergistic workflows for crowdsourcing. In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. Palo Alto, CA, USA: AAAI Press, 2012. p. 2–8. Disponível em: <http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5129>. Acesso em: 1 de outubro de 2015.

LINTOTT, C.; REED, J. Human computation in citizen science. In: MICHELUCCI, P. (Ed.). *Handbook of Human Computation*. New York, USA: Springer, 2013. p. 153–162. ISBN 978-1-4614-8805-7.

LINTOTT, C. J. et al. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 2008. Wiley Online Library, v. 389, n. 3, p. 1179–1189, 2008.

LITTLE, G. et al. Turkit: Human computation algorithms on Mechanical Turk. In: *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 2010. p. 57–66. ISBN 978-1-4503-0271-5.

LIU, P.; LI, Z. Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics*, 2012. v. 42, n. 6, p. 553 – 568, 2012. ISSN 0169-8141.

LOSTAL, E. et al. A case of citizen science for cell biology images analysis. In: *Proceedings of the VII Brazilian e-Science workshop, XXXIII Congresso da Sociedade Brasileira de Computação*. Porto Alegre, Brazil: Brazilian Computer Society, 2013. p. 1855–1862.

LU, J. et al. Hierarchical initialization approach for k-means clustering. *Pattern Recognition Letters*, 2008. v. 29, n. 6, p. 787 – 795, 2008. ISSN 0167-8655.

MAO, A.; KAMAR, E.; HORVITZ, E. Why stop now? predicting worker engagement in online crowdsourcing. In: *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*. Palo Alto, CA, USA: AAAI, 2013. p. 103–111. ISBN 978-1-57735-607-3. Disponível em: <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7498>. Acesso em: 1 de outubro de 2015.

MAO, A. et al. Human computation and multiagent systems: An algorithmic perspective. In: *Proceedings of the twenty-fifth AAAI conference on artificial intelligence*. Palo Alto, CA, USA: AAAI, 2011. p. 1–6.

MARTIN, D. et al. Being a turker. In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. New York, NY, USA: ACM, 2014. p. 224–235.

MASLACH, C.; JACKSON, S. E. The measurement of experienced burnout. *Journal of Organizational Behavior*, 1981. Wiley Online Library, v. 2, n. 2, p. 99–113, 1981.

MASLOW, A. H. A theory of human motivation. *Psychological Review*, 1943. American Psychological Association, v. 50, n. 4, p. 370–396, 1943. ISSN 0033-295X.

MASON, W.; WATTS, D. J. Financial incentives and the "performance of crowds". In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*. New York, NY, USA: ACM, 2009. p. 77–85. ISBN 978-1-60558-672-4.

MCMILLAN, D. W. Sense of community. *J Community Psychol*, 1996. Wiley Online Library, v. 24, n. 4, p. 315–325, 1996.

MEHRZADI, D.; FEITELSON, D. G. On extracting session data from activity logs. In: *Proceedings of the 5th Annual International Systems and Storage Conference*. New York, NY, USA: ACM, 2012. p. 3:1–3:7. ISBN 978-1-4503-1448-0.

MICHALEWICZ, Z.; FOGEL, D. B. *How to Solve It: Modern Heuristics*. 2. ed. Berlin, Germany: Springer, 2004. ISBN 978-3-642-06134-9.

MILLEN, D. R.; PATTERSON, J. F. Stimulating social engagement in a community network. In: *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, 2002. p. 306–313. ISBN 1-58113-560-2.

MORAIS, A.; SANTOS, R.; RADDICK, M. Visualization of citizen science volunteers' behaviors with data from usage logs. *Computing in Science Engineering*, 2015. v. 17, n. 4, p. 42–50, July 2015. ISSN 1521-9615.

MORRIS, M. R. et al. Tweeting is believing?: Understanding microblog credibility perceptions. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, 2012. p. 441–450. ISBN 978-1-4503-1086-4.

MORRIS, R. The emergence of affective crowdsourcing. In: *CHI '11 Workshop on Crowdsourcing and Human Computation*. New York, NY, USA: ACM, 2011. p. 1–4. Disponível em: <http://crowdresearch.org/chi2011-workshop/papers/morris.pdf>. Acesso em: 1 de outubro de 2015.

NORONHA, J. et al. Platemate: Crowdsourcing nutritional analysis from food photographs. In: *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 2011. p. 1–12. ISBN 978-1-4503-0716-1.

O'BRIEN, H. L.; TOMS, E. G. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 2008. Wiley Online Library, v. 59, n. 6, p. 938–955, 2008. ISSN 1532-2890.

PARASURAMAN, R.; JIANG, Y. Individual differences in cognition, affect, and performance: Behavioral, neuroimaging, and molecular genetic approaches. *NeuroImage*, 2012. v. 59, n. 1, p. 70–82, 2012.

PARITOSH, P. Human computation must be reproducible. In: *First International Workshop on Crowdsourcing Web Search*. CEUR-WS.org, 2012. p. 20–25. Disponível em: <http://ceur-ws.org/Vol-842/crowdsearch-paritosh.pdf>. Acesso em: 1 de outubro de 2015.

PINHEIRO, J.; BATES, D. *Mixed-Effects Models in S and S-PLUS*. New Yourk, USA: Springer, 2000. (Statistics and Computing). ISBN 978-0-387-22747-4.

PONCIANO, L.; ANDRADE, N.; BRASILEIRO, F. Bittorrent traffic from a caching perspective. *Journal of the Brazilian Computer Society*, 2013. Springer London, London, v. 19, n. 4, p. 475–491, 2013. ISSN 1678-4804.

PONCIANO, L.; BRASILEIRO, F. *On the Dynamics of Micro- and Macro-task Human Computation Markets*. Universidade Federal de Campina Grande. Laboratório de Sistemas Distribuídos. UFCG-LSD-2013-01, 2013.

PONCIANO, L.; BRASILEIRO, F. Finding volunteers' engagement profiles in human computation for citizen science projects. *Human Computation*, 2014. v. 1, n. 2, p. 245–264, 2014. ISSN 2330-8001.

PONCIANO, L. et al. Considering human aspects on strategies for designing and managing distributed human computation. *Journal of Internet Services and Applications*, 2014. v. 5, n. 1, 2014. ISSN 1869-0238.

PONCIANO, L.; BRASILEIRO, F.; GADELHA, G. Task redundancy strategy based on volunteers' credibility for volunteer thinking projects. In: *AAAI Conference on Human Computation and Crowdsourcing*. Palo Alto, CA, USA: AAAI, 2013. p. 60–61. Disponível em: <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/viewFile/7505/7452>. Acesso em: 1 de outubro de 2015.

PONCIANO, L. et al. Adaptive task replication strategy for human computation. In: *2014 Brazilian Symposium on Computer Networks and Distributed Systems (SBRC)*. Washington, DC, USA: IEEE, 2014. p. 249–257.

PONCIANO, L. et al. Volunteers' engagement in human computation for astronomy projects. *Computing in Science and Engineering*, 2014. IEEE Computer Society, Los Alamitos, CA, USA, v. 16, n. 6, p. 52–59, Nov 2014. ISSN 1521-9615.

QUINN, A. J.; BEDERSON, B. B. *A taxonomy of distributed human computation*. University of Maryland. Human-Computer Interaction Lab. HCIL-2009-23, 2009.

QUINN, A. J.; BEDERSON, B. B. Human computation: A survey and taxonomy of a growing field. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2011. p. 1403–1412. ISBN 978-1-4503-0228-9.

RADDICK, J. et al. Galaxy Zoo: Exploring the motivations of citizen science volunteers. *Astronomy Education Review*, 2010. American Astronomical Society, v. 9, n. 1, p. 010103, 2010. ISSN 15391515.

RADDICK, J. et al. Galaxy Zoo: Motivations of citizen scientists. In: *Bulletin of the American Astronomical Society*. [S.l.]: American Astronomical Society, 2008. v. 40, p. 240.

RAM, N. et al. Cognitive performance inconsistency: Intraindividual change and variability. *Psychol Aging*, 2005. v. 20, n. 4, p. 623–633, 2005. ISSN 0882-7974.

RAO, H.; HUANG, S.; FU, W. What will others choose? how a majority vote reward scheme can improve human computation in a spatial location identification task. In: *AAAI Conference on Human Computation and Crowdsourcing*. [s.n.], 2013. p. 130–137. Disponível em: <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7525>. Acesso em: 1 de outubro de 2015.

RASMUSSEN, J. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, 1983. IEEE, Washington, DC, USA, SMC-13, n. 3, p. 257 –266, may-june 1983. ISSN 0018-9472.

REASON, J. *Human Error*. New York, USA: Cambridge University Press Cambridge, 1990. ISBN 978-0-521-30669-0.

RIEH, S. Y.; DANIELSON, D. R. Credibility: A multidisciplinary framework. *Annual Rev. Info. Sci & Technol.*, 2007. John Wiley & Sons, Inc., New York, NY, USA, v. 41, n. 1, p. 307–364, dez. 2007. ISSN 0066-4200.

ROBSON, C. et al. Comparing the use of social networking and traditional media channels for promoting citizen science. In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, 2013. p. 1463–1468. ISBN 978-1-4503-1331-5.

RODDEN, K.; HUTCHINSON, H.; FU, X. Measuring the user experience on a large scale: User-centered metrics for Web applications. In: *Proceedings of the SIGCHI Conference on*

Human Factors in Computing Systems. New York, NY, USA: ACM, 2010. p. 2395–2398. ISBN 978-1-60558-929-9.

ROGSTADIUS, J. et al. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In: *Fifth International AAAI Conference on Weblogs and Social Media*. Palo Alto, CA, USA: AAAI, 2011. p. 321–328. Disponível em: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2778>. Acesso em: 1 de outubro de 2015.

ROSS, J. et al. Who are the crowdworkers?: Shifting demographics in Mechanical Turk. In: *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2010. p. 2863–2872. ISBN 978-1-60558-930-5.

ROTMAN, D. et al. Dynamic changes in motivation in collaborative citizen-science projects. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM, 2012. p. 217–226. ISBN 978-1-4503-1086-4.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 1987. Elsevier Science Publishers B. V., Amsterdam, Netherlands, v. 20, p. 53–65, 1987.

RZESZOTARSKI, J. M.; KITTUR, A. Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In: *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM, 2011. p. 13–22. ISBN 978-1-4503-0716-1.

SANCHEZ, C. A. et al. Volunteer clouds and citizen cyberscience for LHC physics. *Journal of Physics Conference Series*, 2011. v. 331, n. 6, p. 062022, dez. 2011.

SARMENTA, L. F. G. Sabotage-tolerance mechanisms for volunteer computing systems. *Future Gener. Comput. Syst.*, 2002. Elsevier Science Publishers B. V., Amsterdam, Netherlands, v. 18, n. 4, p. 561–572, mar. 2002. ISSN 0167-739X.

SATZGER, B. et al. Stimulating skill evolution in market-based crowdsourcing. In: RINDERLE-MA, S.; TOUMANI, F.; WOLF, K. (Ed.). *Business Process Management*. Berlin, Germany: Springer, 2011, (Lecture Notes in Computer Science, v. 6896). p. 66–82. ISBN 978-3-642-23058-5.

SAUERMANN, H.; FRANZONI, C. Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences*, 2015. v. 112, n. 3, p. 679–684, 2015.

SAVAGE, N. Gaining wisdom from crowds. *Commun. ACM*, 2012. ACM, New York, NY, USA, v. 55, n. 3, p. 13–15, mar. 2012. ISSN 0001-0782.

SCHALL, D.; SATZGER, B.; PSAIER, H. Crowdsourcing tasks to social networks in bpel4people. *World Wide Web*, 2014. Kluwer Academic Publishers, Hingham, MA, USA, v. 17, n. 1, p. 1–32, jan. 2014. ISSN 1386-145X.

SCHNEIDER, F. B. Implementing fault-tolerant services using the state machine approach: A tutorial. *ACM Comput. Surv.*, 1990. ACM, New York, NY, USA, v. 22, n. 4, p. 299–319, dez. 1990. ISSN 0360-0300.

SCHWARTZ, J. M.; BEGLEY, S. *The Mind and the Brain*. New York, USA: HarperCollins, 2009. ISBN 978-0-06196-198-4.

SCHWARZ, J.; MORRIS, M. Augmenting Web pages and search results to support credibility assessment. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2011. p. 1245–1254. ISBN 978-1-4503-0228-9.

SHANNON, C. E. Prediction and entropy of printed English. *Bell system technical journal*, 1951. Wiley Online Library, v. 30, n. 1, p. 50–64, 1951.

SHENG, V. S.; PROVOST, F.; IPEIROTIS, P. G. Get another label? improving data quality and data mining using multiple, noisy labelers. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2008. p. 614–622. ISBN 978-1-60558-193-4.

SHESHADRI, A.; LEASE, M. SQUARE: A benchmark for research on computing crowd consensus. In: *First AAAI Conference on Human Computation and Crowdsourcing*. Palo Alto, CA, USA: AAAI, 2013. p. 156 – 164. Disponível em: <https://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7550>. Acesso em: 1 de outubro de 2015.

SILVA, D. P. da; CIRNE, W.; BRASILEIRO, F. Trading cycles for information: Using replication to schedule bag-of-tasks applications on computational grids. In: KOSCH, H.; BOSZORMENYI, L.; HELLWAGNER, H. (Ed.). *Euro-Par 2003 Parallel Processing*. Berlin, Germany: Springer, 2003, (Lecture Notes in Computer Science). p. 169–180. ISBN 978-3-540-40788-1.

SIMON, H. A. Theories of bounded rationality. *Decision and organization*, 1972. North-Holland, Amsterdam, v. 1, p. 161–176, 1972.

SIMON, H. A. Invariants of human behavior. *Annual Review of Psychology*, 1990. Annual Reviews, Palo Alto, CA, USA, v. 41, n. 1, p. 1–20, 1990. ISSN 0066-4308.

SIMPSON, M. R. Engagement at work: A review of the literature. *International Journal of Nursing Studies*, 2009. Elsevier Science Publishers B. V., Amsterdam, Netherlands, v. 46, n. 7, p. 1012–1024, 2009. ISSN 0020-7489.

SIMPSON, R.; PAGE, K. R.; ROURE, D. D. Zooniverse: Observing the world's largest citizen science platform. In: *Proceedings of the 23rd International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2014. (WWW '14 Companion), p. 1049–1054. ISBN 978-1-4503-2745-9.

SIMPSON, R. et al. The Milky Way Project first data release: a bubblier galactic disc. *Monthly Notices of the Royal Astronomical Society*, 2012. Wiley Online Library, v. 424, n. 4, p. 2442–2460, 2012.

SINGER, Y.; MITTAL, M. Pricing mechanisms for crowdsourcing markets. In: *Proceedings of the 22nd International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. p. 1157–1166. ISBN 978-1-4503-2035-1.

SINGLA, A.; KRAUSE, A. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In: *Proceedings of the 22nd International Conference on World Wide Web*. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013. p. 1167–1178. ISBN 978-1-4503-2035-1.

SMIRNOV, N. V. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bul. Math. de l'Univ. de Moscou*, 1939. v. 2, p. 3–14, 1939.

SNOW, R. et al. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. p. 254–263.

SPEARMAN, C. The proof and measurement of association between two things. *The American journal of psychology*, 1904. JSTOR, v. 15, n. 1, p. 72–101, 1904.

STAHL, G. Theories of cognition in CSCW. In: BODKER, S. et al. (Ed.). *Proceedings of the 12th European Conference on Computer Supported Cooperative Work*. London, UK: Springer, 2011. p. 193–212. ISBN 978-0-85729-912-3.

STANOVICH, K.; WEST, R. Individual differences in rational thought. *J Exp Psychol Gen*, 1998. American Psychological Association, v. 127, n. 2, p. 161–188, 1998.

STRUYF, A.; HUBERT, M.; ROUSSEEUW, P. Clustering in an object-oriented environment. *Journal of Statistical Software*, 1997. v. 1, n. 4, p. 1–30, 2 1997. ISSN 1548-7660. Disponível em: <http://www.jstatsoft.org/v01/i04>. Acesso em: 1 de outubro de 2015.

SWEDIN, E.; FERRO, D. *Computers: The Life Story of a Technology*. California, USA: Greenwood Press, 2005. (Greenwood technographies). ISBN 9780313331497.

SWELLER, J.; MERRIENBOER, J. J. G. V.; PAAS, F. G. W. C. Cognitive architecture and instructional design. *Educ Psychol Rev*, 1998. Kluwer Academic Publishers-Plenum Publishers, v. 10, p. 251–296, 1998. ISSN 1040-726X.

TANENBAUM, A. S.; STEEN, M. v. *Distributed Systems: Principles and Paradigms (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006. ISBN 0-132-39227-5.

TURING, A. M. Computing machinery and intelligence. *Mind*, 1950. JSTOR, p. 433–460, 1950.

VAROL, O. et al. Evolution of online user behavior during a social upheaval. In: *Proceedings of the 2014 ACM Conference on Web Science*. New York, NY, USA: ACM, 2014. p. 81–90. ISBN 978-1-4503-2622-3.

WANG, D. et al. Recursive factfinding: A streaming approach to truth estimation in crowdsourcing applications. In: *IEEE 33rd International Conference on Distributed Computing Systems*. Washington, DC, USA: IEEE Computer Society, 2013. p. 530–539. ISBN 978-0-7695-5000-8.

WATHEN, C. N.; BUREL, J. Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology*, 2002. John Wiley & Sons, Inc., New York, NY, USA, v. 53, n. 2, p. 134–144, jan. 2002. ISSN 1532-2882.

WEN, M.; ROSE, C. P. Understanding participant behavior trajectories in online health support groups using automatic extraction methods. In: *Proceedings of the 17th ACM International Conference on Supporting Group Work*. New York, NY, USA: ACM, 2012. p. 179–188. ISBN 978-1-4503-1486-2.

WHITEHILL, J. et al. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In: *Advances in Neural Information Processing Systems*. NY, USA: Curran Associates, Inc., 2009. v. 22, p. 2035–2043. ISBN 9781615679119.

WIGGINS, A.; CROWSTON, K. Goals and tasks: Two typologies of citizen science projects. In: *Proceedings of the 45th Hawaii International Conference on System Sciences*. Los Alamitos, CA, USA: IEEE Computer Society, 2012. p. 3426–3435.

WITKOWSKI, J. et al. Dwelling on the negative: Incentivizing effort in peer prediction. In: *AAAI Conference on Human Computation and Crowdsourcing*. Palo Alto, CA, USA: AAAI, 2013. p. 190–197. Disponível em: <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7558>. Acesso em: 1 de outubro de 2015.

YI, J. et al. Inferring users' preferences from crowdsourced pairwise comparisons: A matrix completion approach. In: *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*. Palo Alto, CA, USA: AAAI, 2013. p. 207–215. ISBN 978-1-57735-607-3. Disponível em: <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/viewFile/7536/7421>. Acesso em: 1 de outubro de 2015.

YUEN, M.-C.; KING, I.; LEUNG, K.-S. A survey of crowdsourcing systems. In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT)*. Washington, DC, USA: IEEE, 2011. p. 766 –773.

ZHANG, H. et al. Human computation tasks with global constraints. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2012. p. 217–226. ISBN 978-1-4503-1015-4.

ZHANG, J. et al. Sentiment bias detection in support of news credibility judgment. In: *Proceedings of the 44th Hawaii International Conference on System Sciences*. Washington, DC, USA: IEEE, 2011. p. 1–10. ISSN 1530-1605.

Apêndice A

Computação Antes dos Computadores Digitais

History doesn't repeat itself, but it does rhyme.

Mark Twain

"A história não se repete, mas rima." Essa frase remete à ideia de que o presente não é igual ao passado, mas, certamente, ele tem consigo elementos que se remetem ao passado. Neste trabalho, essa ideia se mostra presente quando se analisa a história do uso do poder cognitivo de seres humanos para realizar computações e o próprio uso do termo "computador".

O uso do poder cognitivo de seres humanos para realizar computações não é um conceito novo. O termo "computador", que atualmente é utilizado para designar máquinas de computar, até a primeira metade do século XX era usado para designar seres humanos que tinham como atividade profissional realizar computações (CERUZZI, 1991; GRIER, 2007). Computadores eram seres humanos que trabalhavam fazendo cálculos matemáticos. Os relatos mais antigos dessa atividade remontam ao século XVI. Ela pode ser analisada considerando dois períodos históricos. No primeiro período, computação por humanos era uma atividade de baixa escala, realizada em pequenas equipes. No segundo período há uma expansão na demanda de computação, o que motiva o surgimento de grandes organizações dedicadas a essa atividade.

Enquanto uma atividade baixa escala, computação por humanos esteve amplamente relacionada à análise matemática e à pesquisa científica (GRIER, 2007). Era geralmente realizada por um cientista sozinho ou em conjunto com seus estudantes e auxiliares. Essa atividade remonta ao início do desenvolvimento da aritmética logarítmica por Henry Briaggs (1561-1630), da trajetória balística por Galileu Galilei (1564-1642), da computação planetária nas tabelas de Rudolphine (*Rudolphine Tables*) por Johannes Kepler (1571-1630) e a invenção do cálculo por Isaac Newton (1642-1727) e Gottfried Wilhelm Leibniz (1646-1716) nos anos de 1680. Nesta fase, surgiram os primeiros conceitos de organização da atividade de computação. O principal conceito é o de divisão de trabalho introduzido pelos astrônomos Alexis-Claude Clairaut (1713-1765), Joseph-Jérôme de Lalande (1732-1807) e Nicole-Reine Aplaude (1723-1788) quando realizavam as computações para prever o retorno do cometa Halley. Para otimizar o trabalho, identificar e tratar erros, eles criaram os princípios de dividir uma grande computação em pequenos pedaços, reunir os resultados parciais em um resultado final e checar os resultados.

Nesse período histórico, computação por humanos se tornou uma atividade de larga escala em razão do crescimento da demanda e da necessidade de maior desempenho (SWEDIN; FERRO, 2005; GRIER, 2007). O uso de computação por humanos se expandiu para além de atividades científicas e surgiram problemas mais complexos que demandaram maior poder computacional para serem solucionados. Por exemplo, em 1790, Napoleão requisitou a produção de novas tabelas decimais que ficou conhecido como o maior projeto de criação de tabelas no mundo. O responsável por esse projeto foi Gaspard de Prony (1755-1839). Para executar esse projeto, ele criou a primeira grande organização dedicada a realizar computação. Essa organização era uma fábrica que empregava 96 pessoas para produzir 7.000 computações por dia. A organização da fábrica era inspirada nos princípios de divisão do trabalho de Adam Smith (1723-1790). De uma forma geral, a estrutura das organizações de computação foi baseada em teorias que estavam em evidência naquele período histórico. Tais como, divisão do trabalho, produção em massa e administração. Esse tipo de computação foi amplamente utilizado até a primeira metade do século XX. Nos últimos anos, ela era utilizada principalmente na NASA[1], no CERN[2] e em órgãos governamentais.

[1]Um relato da NASA é apresentado em http://crgis.ndc.nasa.gov/historic/Human_Computers. Último acesso em 01 de setembro de 2015.

[2]Um relato do CERN é apresentado em http://timeline.web.cern.ch/timelines/Computing-at-CERN. Último

Na NASA e outros órgãos do governo dos Estados Unidos, a maior parte das pessoas que atuavam como computadores eram do sexo feminino (LIGHT, 1999). Predominavam ex-professoras de ensino médio com graduação em matemática. A maior parte dos trabalhos realizados eram transcrever dados existentes em filmes, fazer cálculos e plotar dados. Segundo a Work Project Administration (WPA) dos Estados Unidos, computadores trabalhavam 32 horas por semana. De uma forma geral, computação era vista como uma subprofissão. Entretanto, havia pessoas com grande reputação e reconhecimento por habilidade de realizar cálculos. Por exemplo, Wim Klein[3](1912-1986) ficou conhecido como o primeiro supercomputador do CERN. Ele se destacava por uma grande habilidade de fazer cálculos extremamente complexos de forma rápida e precisa.

Essa primeira fase de computação por humanos deixou um importante legado para o que veio a se tornar a Ciência da Computação. A organização das fábricas de computar, os padrões de erros observados e os mecanismos desenvolvidos para identificar e tratar esses erros inspiraram Charles Babbage (1791-1871) na proposta da primeira máquina de calcular (*Difference Engine*) por volta de 1822. O próprio conceito de "computação" definido por Alan Turing (1912-1954) é inspirado na forma como os computadores humanos realizavam as computações (TURING, 1950).

acesso em 01 de setembro de 2015.

[3]O primeiro computador do CERN http://timeline.web.cern.ch/wim-klein-cerns-first-computer. Último acesso em 01 de setembro de 2015.

Apêndice B

A Expressão ‘Computação por Humanos’ e Expressões Correlatas

The beginning of wisdom is the definition of terms.

Sócrates

“O começo da sabedoria é a definição de termos.” Essa frase atribuída ao filósofo Sócrates[1] destaca que para que alguma sabedoria possa ser construída em uma argumentação é importante uma definição precisa dos termos envolvidos na argumentação. Em essência, para que uma argumentação seja conduzida, é necessário que os termos nela tratados sejam entendidos e aceitos pelos envolvidos. Quando se trata de uma argumentação em uma área nova ou pouco conhecida, a definição de termos se torna fundamental. Neste apêndice, justifica-se o uso da expressão ‘computação por humanos’ e algumas traduções e termos utilizados neste documento.

O termo “computação” pode ser definido como a ação de mapear uma entrada em uma saída por meio do processamento de um conjunto finito de instruções (TURING, 1950). Esse mapeamento pode ser realizado por uma máquina ou por um ser humano (TURING, 1950; AHN, 2005; QUINN; BEDERSON, 2011; LAW, 2011). Em inglês, utiliza-se a expressão *human*

[1]Não se pode confirmar muitas das frases atribuídas a Sócrates, pois ele não deixou documentos escritos. Entretanto, o conteúdo dessa frase se assemelha ao pensamento filosófico conduzido por ele e seguido por seus seguidores como Platão e Xenofonte. Para uma discussão mais sobre isso, acesse http://www.askphilosophers.org/question/5187. Último acesso em 01 de setembro de 2015.

computation para denotar o modelo de computação no qual a ação de computar é realizada por um ser humano. A expressão *machine computation*, por sua vez, designa o modelo de computação no qual a ação de computar é realizada por uma máquina. Nesse idioma, constrói-se, portanto, o paralelo *human computation* e *machine computation*.

Em português, a expressão *human computation* pode ser traduzida literalmente como "computação humana". Entretanto, essa tradução literal gera alguns problemas que convém serem explicitados e evitados. Na expressão "computação humana", o termo "humana" é empregado como um adjetivo que qualifica a computação. Esse adjetivo é geralmente entendido na língua portuguesa como sinônimo de "bondoso, benfazejo, compassivo" (FERREIRA; ANJOS, 1988). Isso pode erroneamente remeter o leitor à ideia de que "computação humana" se refere a uma computação que é caracterizada por ser bondosa ou que trata de aspectos humanitários. Além disso, ao se traduzir literalmente *human computation* como "computação humana", não é claro qual seria a tradução literal da expressão *machine computation*.

Optou-se por uma tradução contextual com o propósito de facilitar a compreensão e evitar ambiguidades. Traduz-se a expressão *human computation* como "computação por humanos". A expressão *machine computation*, por sua vez, é traduzida como "computação por máquinas". Nessas traduções, a preposição "por" é utilizada para indicar quem realiza a ação de computar. Assim, tem-se o paralelo "computação por humanos" e "computação por máquinas" que explicita claramente a principal diferença entre esses modelos de computação.

Adicionalmente, é importante ressaltar que também na literatura em inglês ainda não existe consenso sobre a expressão mais adequada. Existem autores que evitam o emprego da expressão *human computation*, preferindo expressões que caracterizam mais diretamente o modelo de computação, como: *human-based computation* – usada para designar o modelo de computação no qual a ação de computar é realizada por um ser humano – e *machine-based computation* – usada para designar o modelo de computação no qual a ação de computar é realizada por uma máquina.

Apêndice C

Trabalhos Futuros com Evidências

"Once we accept our limits, we go beyond them"

Albert Einstein

"Uma vez que nós aceitamos nossos limites, vamos além deles." Essa frase remete à existência de limites e ao reconhecimento deles como requisito para se criar condições para que eles possam ser superados. Essa frase é especialmente importante no contexto de pesquisas científicas, onde, em sua essência, limites são compreendidos, definidos e avançados progressivamente. Os limites de um trabalho podem surgir de diversos fatores como a complexidade de se tratar todo o ambiente de estudo por limitação dos recursos disponíveis, como dados, poder computacional e, até mesmo, tempo. Quando se estabelece e se aceita os limites, tem-se com clareza o escopo tratado e a fronteira que pode ser expandida por trabalhos futuros.

A pesquisa descrita neste documento foi exploratória como uma das primeiras a investigar o engajamento e a credibilidade de trabalhadores em sistemas de computação por humanos. Como toda pesquisa exploratória, ela tratou os aspectos mais simples que têm potencial de seres relevantes. Naturalmente, a partir da construção apresentada neste trabalho, diversos outros estudos no contexto de engajamento, credibilidade e replicação de tarefas podem ser conduzidos. Um contexto de especial interesse são os sistemas com múltiplos projetos.

Ao se criar um sistema de computação por humanos exclusivamente para executar as tarefas de um projeto, tem-se diversos custos operacionais. Além disso, é necessário condu-

zir todo um processo de recrutamento de trabalhadores para o sistema. De outro modo, em teoria, quando se hospeda o projeto em um sistema existente como Zooniverse.org e Crowdcrafting.org, obtém-se uma redução dos custos operacionais e beneficia-se dos trabalhadores que já contribuem executando tarefas nesses sistemas (FORTSON et al., 2012). Esse é uma das razões da existência de sistemas com múltiplos projetos. Nesses sistemas, o engajamento e credibilidade dos trabalhadores ao longo dos diversos projetos em que eles atuam é um comportamento de interesse. Há pelo menos duas perspectivas imediatas, uma perspectiva dos mantenedores do sistema no qual os projetos são hospedados e outra perspectiva dos usuários que criam projetos no sistema.

Os mantenedores do sistema têm interesse (e até mesmo uma necessidade) de que sejam criados no sistema novos projeto que sejam de interesse dos trabalhadores. Projetos assim têm potencial de manter engajados os trabalhadores que já atuam no sistema e também de atrair (recrutar) novos trabalhadores para o sistema. Os usuários, por sua vez, ao criarem um projeto no sistema, têm interesse de se beneficiar de trabalhadores que já atuam no sistema. Se alguns desses trabalhadores atuarem no novo projeto, tem-se uma redução do custo/esforço no recrutamento de trabalhadores. No entanto, essas são considerações que, até onde se sabe, nunca foram medidas em sistemas reais.

A seguir são destacados dois estudos nessa direção: atração de trabalhadores em sistemas com múltiplos projetos (Seção C.1) e engajamento em sistemas multi-projeto (Seção C.2). Ambos os estudos utilizam dados do sistema Crowcrafting. Foram coletados dados de 242 projetos hospedados no sistema no período entre 07/07/2012 e 28/03/2014 (Tabela C.1). As descrições e as tarefas de todos os projetos foram analisadas e contatou-se que apenas 22 (8%) deles eram projetos reais, os demais eram projetos de demonstração ou projetos de teste. Apenas os projetos reais foram analisados.

Tabela C.1: Resumo estatístico da base de dados do sistema Crowdcrafting.

	Todos os projetos	**Projetos selecionados**
Número de projetos	242	22
Número de trabalhadores	28.609	26.113
Número de tarefas	1.380.300	1.252.502
Número de trabalhadores regulares	2.063	1.697
Número de trabalhadores que atuaram em múltiplos projetos	4.592	3.452

C.1 Atração de Trabalhadores em Sistemas com Múltiplos Projetos

É intuitivo se perguntar em que medida os usuários que criam projetos no sistema realmente recrutam novos trabalhadores para o sistema e em que medida eles herdam trabalhadores que já atuavam em outros projetos no sistema. Visando responder essa pergunta, para cada projeto, calculou-se o número de trabalhadores herdados n_i e o número de trabalhadores recrutados n_r. Então, mediu-se a "diferença relativa de trabalhadores herdados e recrutados", definido dela fórmula $(n_i - n_r)/min(n_i, n_r)$. Quando o resultado é 0, a quantidade de trabalhadores recrutados é igual à quantidade de trabalhadores herdados. Quando o resultado é positivo, ele diz em quantas vezes a quantidade de trabalhadores herdados é maior do que a quantidade de trabalhadores recrutados. Quando o valor for negativo, ele diz em quantas vezes a quantidade de trabalhadores recrutados é maior do que a quantidade de trabalhadores herdados.

A Figura C.1 mostra a distribuição dos valores no conjunto de 22 projetos selecionados. A maioria dos projetos (cerca de 72%) recruta uma quantidade de trabalhadores maior que a quantidade de trabalhadores que eles herdam da plataforma. O número de trabalhadores recrutados pode ser mais do que 300 vezes maior do que a quantidade de trabalhadores herdados. Isso indica que a maioria dos projetos recruta mais trabalhadores do que herdam do sistema.

C.2 Engajamento de Trabalhadores em Sistemas com Múltiplos Projetos

É de interesse saber, em que medida os trabalhadores recrutados pelo usuário são mais engajados que os trabalhadores herdados do sistema. Para responder essa pergunta, para cada projeto, calculou-se a média das tarefas desempenhadas por trabalhadores herdados t_i e a média das tarefas desempenhadas por trabalhadores recrutados t_r. Então, mediu-se a "diferença relativa na média de tarefa executadas por trabalhadores herdados e recrutados.", definida como $(t_i - t_r)/min(t_i, t_r)$. Quando o valor é 0, a média de tarefas realizadas pelos

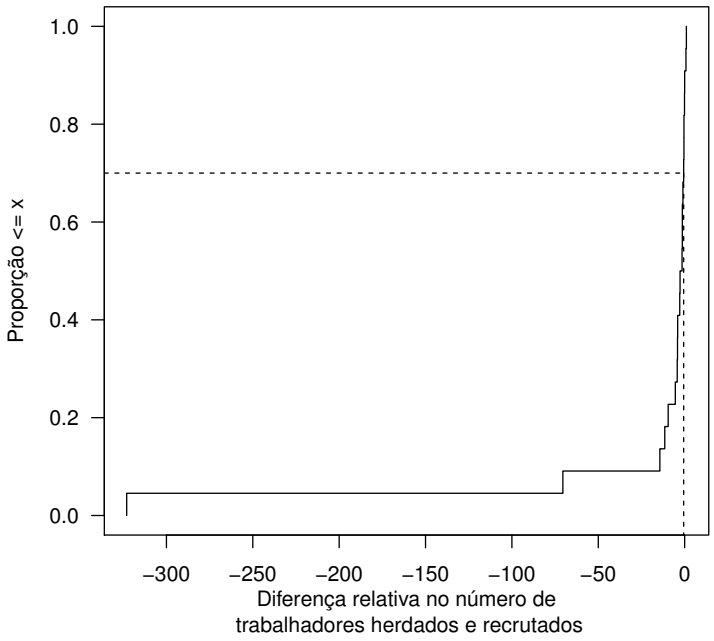


Figura C.1: Função de distribuição acumulada (FDA) da diferença relativa no número de trabalhadores herdados do sistema e do número de trabalhadores recrutados pelos usuários que criaram os projetos.

trabalhadores recrutados é igual à média de tarefas realizadas pelos trabalhadores herdados. Quando o valor é positivo, ele diz quantas vezes a média ou tarefas executadas por trabalhadores herdados é maior do que a quantidade de trabalhadores recrutados. Quando o valor for negativo, ele diz quantas vezes a média de tarefa realizada por trabalhadores recrutados é maior do que a média da tarefa realizada por trabalhadores herdados.

A Figura C.2 mostra a distribuição dos valores no conjunto de 22 projetos selecionados. Na maioria dos projetos (cerca de 68%), a quantidade média de tarefas executadas por trabalhadores herdados é maior que a quantidade média de tarefas executadas por trabalhadores recrutados pelo usuário. Trabalhadores herdados podem executar uma quantidade de tarefas até 30 vezes maior que os trabalhadores recrutados.

Conclusão - Novos projetos herdam poucos trabalhadores da plataforma, mas esses trabalhadores herdados são mais engajados do que aqueles que foram recrutados pelo usuário. Isso revela um comportamento esperado, pois recrutamento nesse tipo de projeto geralmente consiste em campanhas de divulgação em massa (*broadcasting*), por exemplo, em redes so-

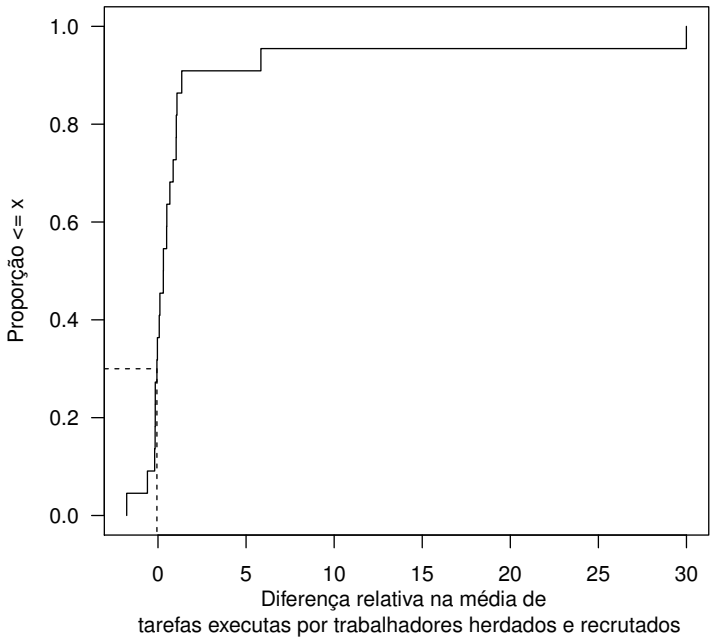


Figura C.2: Função de distribuição acumulada (FDA) da diferença relativa no número de tarefas executadas por trabalhadores herdados do sistema e do número de tarefas executadas por trabalhadores recrutados pelos usuários que criaram os projetos.

ciais (ROBSON et al., 2013). Essas campanhas têm um potencial de atrair muitos trabalhadores que são apenas curiosos e que não pretendem ter um longo engajamento (PONCIANO et al., 2014b). Os trabalhadores que são herdados de outros projetos na plataforma, já apresentam uma tendência a natural a serem mais engajados. Isso explica o fato de que em média eles executam mais tarefas.

A maior deficiência desses estudos se refere aos dados utilizados. Muitos projetos tiveram de ser removidos da base de dados, pois eram apenas testes ou demonstrações. No momento, não existe disponível dados que permitam realizar esse estudo com mais rigor. Em comunicação pessoal, os mantenedores do sistema Crowdcrafting informaram que estão em vias de resolver esse problema. Uma nova alternativa que surge e que permitirá a condução desse tipo de estudo é o sistema Contribua (https://contribua.org/), um sistema multiprojeto com dados abertos.