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

# Perfis de contribuidores em sites de perguntas e respostas

Adabriand Andrade Furtado

Dissertação submetida à Coordenação do Curso 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 Mestre em Ciência da Computação.

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

Nazareno Ferreira de Andrade
(Orientador)

Campina Grande, Paraíba, Brasil

**FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG**

F992p Furtado, Adabriand Andrade.

Perfis de contribuidores em sites de perguntas e respostas / Adabriand Andrade Furtado. – Campina Grande, 2013.

67 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2013.

"Orientação: Prof. Dr. Nazareno Ferreira de Andrade."

Referências.

1. Sites de Perguntas e Respostas. 2. Mineração de Dados e Aprendizagem de Máquina. 3. Estudos da WEB. I. Andrade, Nazareno Ferreira de. II. Título.

CDU 004.773.7(043)

# Resumo

Sites de perguntas e respostas (Q&A) têm se mostrado um recurso valioso em ajudar pessoas a resolverem seus problemas do dia-a-dia. Esses sites atualmente permitem que um grande número de contribuidores troque conhecimento por meio de atividades como criar perguntas, respostas, comentários e avaliar o conteúdo gerado. Dado o tamanho desses sites, é natural que seus usuários contribuam em diversas quantidades e criem conteúdo de qualidade variada. Interessado na diversidade de comportamentos, nosso trabalho avança o presente conhecimento sobre sites de Q&A realizando uma análise multifacetada dos contribuidores, que considera indicadores de participação e habilidade, para identificar perfis de comportamento típicos nesses sites. Examinando toda a atividade dos contribuidores de 36 sites de Q&A da plataforma Stack Exchange, nossa análise revela nove perfis comportamentais que agrupam os usuários de acordo com a quantidade e qualidade de suas contribuições. Baseado nesses perfis, analisamos a composição dessas comunidades e a importância dos grupos formados pelos perfis na produção de conteúdo. Essas análises sugerem que diferentes composições e produções dos grupos são consequências do tamanho e tipo de tópico discutido no site. Adicionalmente a essas análises, conduzimos um estudo longitudinal no maior dos 36 sites, Super User, visando investigar o aspecto de dinâmica do comportamento dos contribuidores. Nesse estudo analisamos a evolução da composição do site, mudanças típicas de perfis, e a relação entre a experiência e o perfil assumido pelo contribuidor. Observamos que embora os usuários mudem de perfil com certa frequência e determinados perfis demandem experiência, a composição do Super User é notavelmente estável ao longo do tempo.

# Abstract

Question-and-answer (Q&A) sites have shown to be a valuable resource for helping people to solve their everyday problems. These sites currently enable a large number of contributors to exchange expertise through activities like creating questions, answers or comments and assessing the created content. Given the size of these sites, it is natural that their users contribute in diverse amounts and create content of varying quality. Concerned with diversity of behaviors, our work advances present knowledge about Q&A sites conducting a multifaceted analysis of contributors, that accounts their participation and ability, to identify typical behavioral profiles in these sites. By examining all contributors' activity from 36 Q&A site in Stack Exchange platform, our analysis unveiled nine behavioral profiles that group users according to the quality and quantity of their contributions. Based on these profiles, we analyze the communities' composition and the importance of the groups formed by these profiles in the content production. These analyzes suggest that different compositions and group's production are consequences of the size and type of the topic discussed in the site. In addition to these analyzes, we conduct a longitudinal study in the largest of 36 sites, Super User, aiming to investigate the dynamic aspect of the contributors' behavior. In this study, we analyze changes in the community's composition, common profile transitions, and the relation between experience and profile assumed by the contributor. We observe that although users change profiles with some frequency and certain profiles demand experience, the Super User's composition is remarkably stable over time.

# Agradecimentos

Começo agradecendo a minha família, a meus pais Marli Andrade e Antonio Furtado (*in memoriam*) e a minhas irmãs Alessandra Furtado e Amanda Furtado, o apoio e incentivo recebido durante o desenvolvimento do meu trabalho de mestrado. Estendo esses agradecimentos aos meus grandes amigos – Gustavo Jansen, Lígia Saraiva, Mariana Vasconcelos, Paloma Vasconcelos, Juliana Vasconcelos, Ícaro Bolconte, Lucas Azevedo e Renato Miceli – que contribuíram em aliviar o stress do trabalho em diversos momentos.

Além deles, agradeço fortemente a todos os colegas e professores que deram contribuições importantíssimas para este trabalho. Especialmente, agradeço ao meu orientador Nazareno o apoio e as boas discussões que contribuíram para melhorar a qualidade técnica do trabalho; e ao co-orientador não oficial, Fubica (Francisco Brasileiro), que ajudou a guiar este trabalho desde os estágios iniciais com excelentes revisões.

Não posso deixar de agradecer aos colegas do grupo de pesquisa em sistemas colaborativos, que participaram desta pesquisa de forma direta ou indireta: Nigini Oliveira, Lesandro Ponciano, Aline Marques, Andryw Marques, José Farias, Jeymisson Oliveira, Milena Araújo e Renata Saraiva.

Também agradeço aos colegas do LSD e professores do DSC que propiciaram excelentes discussões, em conversas no corredor e em apresentações deste trabalho: Leandro Balby, Raquel Lopes, Lívia Sampaio, Andrey Brito, Thiago Emmanuel, Marcus Carvalho, Abmar Granjeiro, Fábio Jorge, Giovanni Farias, Rodrigo Duarte, Ricardo Araújo e David Lino.

Meus agradecimentos também vão para a equipe de suporte do LSD e da COPIN: a Rúbia Ramos e Rebeka Lemos, pela agilidade em atender e resolver nossos pedidos; a Elayne Leal e Josicleide Souza, por manter a infra-estrutura do laboratório funcionando em perfeitas condições.

Por fim, agradeço aos programas CAPES e CNPq o apoio financeiro que possibilitou a execução deste trabalho de mestrado.

# Conteúdo

# Lista de Símbolos

Q&A - *Question and answer*

SE - *Stack Exchange*

SO - *StackOverflow*

YA - *Yahoo! Answers*

Naver KiN - *Naver Knowledge-iN*

FAQ - *Frequently asked questions*

UMPerguntas - *Utilidade média das perguntas*

UMRespostas - *Utilidade média das respostas*

UMComentários - *Utilidade média dos comentários*

AC - *Estado de abandono da comunidade*

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

Sites de perguntas e respostas (Q&A) têm possibilitado o trabalho colaborativo de milhares de pessoas na busca por soluções para seus problemas. Nesses sites, usuários trocam conhecimento através de atividades como criar questões, respostas, comentários e votar na qualidade desses, possibilitando encontrar conteúdo útil.

Atualmente, sites como Yahoo! Answers, Naver Knowledge-iN e StackOverflow têm mostrado potencial em atrair um grande volume de contribuidores voluntários para criar bases dinâmicas e valiosas de conhecimento [2; 22; 25]. Dados do Yahoo! Answers mostram que o site chegou à marca de 24 milhões de questões resolvidas [2], enquanto que dados do Naver Knowledge-iN apontam que a base de conhecimento do site agrega mais de 60 milhões de postagens (perguntas e respostas) geradas por seus usuários [25]. Já a comunidade StackOverflow, com aproximadamente 2 milhões de contribuidores [8], se destaca pelo seu alto desempenho. As perguntas nesse site são respondidas em um tempo mediano de aproximadamente 11 minutos [22].

Como efeito do tamanho considerável dessas comunidades, é de se esperar que contribuidores em sites de Q&A exibam comportamentos diversos na criação de conteúdo – seja dando preferência à criação de um determinado tipo de contribuição (e.g. somente respostas) ou demostrando alguma habilidade em criar conteúdo. Trabalhos nesta linha têm mostrado que os usuários possuem diferentes motivações para contribuir [25; 11; 33], passam diferentes quantidades de tempo contribuindo [22; 25], demonstram habilidades diversas [28; 26], e por fim, apresentam preferências diversas em criar diferentes tipos de conteúdo nesses sites [2; 32].

Na perspectiva de entender os variados comportamentos dos contribuidores, diversos trabalhos examinaram dados históricos de contribuição dos usuários para identificar perfis típicos de comportamento em comunidades de produção coletiva. Esses estudos abrangem comunidades como *newsgroup* (e.g. Usenet [37; 34; 9; 15; 38]), Wikipedia [36; 6; 21], comunidades de desenvolvimento de *software open-source* [24], comunidades de compartilhamento de conteúdo [10; 5], sites de recomendação de filmes [11], e também sites de Q&A (e.g. Yahoo! Answers [2], Naver Knowledge-iN [25] e StackOverflow [22]).

Considerando esses trabalhos, nosso estudo expande o presente conhecimento sobre o comportamento de contribuidores em sites de Q&A ao responder as seguintes questões de pesquisa:

**I - Relacionadas aos comportamentos típicos dos contribuidores de sites de Q&A.**

1. Que perfis de comportamento são identificados considerando as perspectivas de quantidade e qualidade das contribuições dos usuários?
2. Como o comportamento dos usuários segundo esses perfis muda ao longo do tempo?

**II - Relacionadas à organização estrutural das comunidades de sites de Q&A.**

1. Qual a composição dessas comunidades segundo esses perfis?
2. Como a composição dessas comunidades evolui ao longo do tempo?
3. Qual o papel dos diferentes grupos de contribuidores formados pelos perfis na produção de conteúdo?

O primeiro grupo de questões de pesquisa visa analisar a diversidade de comportamento dos contribuidores em sites de Q&A. Especificamente, nossa primeira questão objetiva identificar perfis típicos de comportamento dos contribuidores levando em consideração dimensões de quantidade (indicador de participação) e qualidade (indicador de habilidade) de suas contribuições. Adicionalmente, a segunda questão desse grupo foca em entender o comportamento dinâmico dos contribuidores – sua mudança de comportamento ao longo do tempo. Aliada a essas análises, o segundo grupo de questões objetiva examinar aspectos estruturais de sites de Q&A, tais como a composição de sua comunidade e evolução ao longo do tempo, e a importância dos grupos de contribuidores de diferentes perfis na produção de conteúdo.

Entender como cada tipo de contribuidor colabora para o funcionamento desses sites ajuda a melhorar o gerenciamento dos mesmos. Administradores podem usar esse conhecimento no desenvolvimento de estratégias para promover ou inibir certos comportamentos na comunidade. Além disso, identificar as diferentes habilidades dos contribuidores pode auxiliar o desenvolvimento de mecanismos de alocação de tarefas no site. Tarefas nesse contexto podem ser entendidas como uma sugestão de questão para ser respondida, ou de revisão de algum conteúdo que necessite de moderação. Por fim, a análise da diversidade de comportamento, em geral, ajuda a criar teorias sobre o comportamento das pessoas em grandes grupos colaborativos online.

Este trabalho contribui para o entendimento sobre a produção coletiva de conhecimento em sites de Q&A examinando os comportamentos típicos dos contribuidores. Para tal, utilizamos dados históricos de 36 sites da plataforma de Q&A Stack Exchange (SE), e extraímos um conjunto de métricas que descrevem quantidade e a qualidade de suas contribuições em um dado período de tempo. Usando esses dados, aplicamos a análise de agrupamento para identificar grupos de contribuidores com comportamento semelhante em perspectivas de longo e curto prazo.

Na perspectiva de longo prazo – usando dados de todo o período de atividade dos usuários – investigamos os perfis de comportamento típicos dos contribuidores nos 36 sites da plataforma SE. Baseado nos perfis identificados, utilizamos a classificação resultante dos contribuidores para analisar as diferentes composições das comunidades e a produção de conteúdo dos grupos de usuários formados pelos perfis.

Na perspectiva de curto prazo – usando dados de atividade em janelas pequenas de tempo – realizamos um estudo longitudinal no maior dos 36 sites, Super User. Nesse experimento investigamos os perfis de comportamento de curto prazo dos contribuidores, a evolução da composição da comunidade Super User e a dinâmica de mudança de comportamento dos usuários desse site.

Ambas as análises de agrupamento nas perspectivas de longo e curto prazo revelam nove perfis comportamentais, os quais podem ser resumidos em quatro tipos: (I) contribuidores *sem habilidade marcante*, usuários de baixa a média atividade e habilidade; (II) *imperitos*, usuários com contribuições avaliadas como pouco úteis; (III) *experts*, contribuidores hábeis em realizar um tipo de atividade; e (IV) *ativistas*, contribuidores de alta atividade.

A análise dos perfis contribui para o gerenciamento desses sites chamando a atenção, por exemplo, para a necessidade de orientar contribuidores imperitos. Além disso, essa análise mostra que contribuidores de alta atividade e contribuidores de alta habilidade formam grupos distintos. A separação desses grupos pode apoiar a concepção de mecanismos de alocação de tarefas na comunidade.

A análise de composição e de produção nos 36 sites revela que esses sites, em geral, são mantidos na sua maioria por uma massa de contribuidores sem habilidade marcante e que atuam esporadicamente, e por um pequeno grupo de contribuidores de alta atividade. Ambas as análises também revelam variações notáveis do retrato geral das composições e produções dos grupos. Nosso estudo indica que essas variações estão correlacionadas com o tamanho da comunidade e o tipo de tópico discutido no site.

Na análise de perfis de curto prazo, a investigação da evolução da composição do Super User mostra uma notável estabilidade da distribuição dos usuários nos perfis ao longo do tempo. Contudo, seus contribuidores mudam de perfil com certa frequência. essa análise revela que a maioria dos contribuidores, exceto ativistas, tende a mudar para perfis de menor atividade e a abandonar o site.

Examinando a relação entre os perfis e a experiência do contribuidor (tempo ativo no site), identificamos que ativistas e experts são perfis associados a contribuidores experientes, enquanto imperitos ocorrem em maior frequência entre novatos. Observar uma relativa alta proporção de imperitos entre novatos chama atenção para a necessidade de orientar novos usuários principalmente em suas primeiras contribuições.

## 1.1 Organização do documento

O restante deste documento está organizado da seguinte forma. No Capítulo 2, posicionamos este estudo dentro dos trabalhos relacionados e apresentamos nossa contribuição. Em seguida, no Capítulo 3, detalhamos o modelo básico de funcionamento de sites de Q&A e a plataforma Stack Exchange.

Depois de apresentado o SE, descrevemos no Capítulo 4 a análise de perfis de comportamento de longo prazo em 36 sites da plataforma. Nesse capítulo apresentamos os dados utilizados, o método de agrupamento para identificar os perfis e a interpretação dos resulta-

dos obtidos.

Utilizando os perfis resultantes da análise de agrupamento nos 36 sites, caracterizamos a composição dos sites no Capítulo 5 e a produção dos perfis no Capítulo 6. Nesses capítulos descrevemos o aspecto geral dos sites, sites com configurações particulares e também discutimos implicações dessas análises.

Na segunda etapa deste trabalho, Capítulo 7, apresentamos nosso estudo sobre a dinâmica de comportamento dos contribuidores do site Super User. Esse capítulo apresenta a caracterização de perfis de curto prazo, a evolução da composição do Super User, as mudanças típicas de comportamento dos contribuidores ao longo do tempo, e a relação entre a experiência e o perfil assumido pelo contribuidor.

No Capítulo 8, sintetizamos toda a nossa caracterização do comportamento de contribuidores em sites de Q&A, indicando os resultados principais e suas implicações na melhoria desses sites. Finalmente, discutimos as limitações desse estudo e as possíveis direções para trabalhos futuros.

# Capítulo 2

# Perfis de contribuidores em comunidades online

Grande parte dos trabalhos sobre perfis de contribuidores em comunidades online buscaram identificar papéis típicos assumidos pelos usuários e/ou examinar o comportamento de contribuidores que assumem um perfil conhecido *a priori*, tais como moderadores e *lurkers*. Em ambos os casos, essas análises têm como objetivo ajudar na interpretação do comportamento coletivo e auxiliar o gerenciamento desses sites, apoiando o desenvolvimento de estratégias de alocação de tarefas e de mecanismos para promover ou inibir determinados comportamentos.

## 2.1 Caracterizando o comportamento dos contribuidores

Diversos trabalhos examinaram dados históricos, ou conduziram trabalhos de campo, para identificar perfis de contribuidores salientes (ou papéis) em comunidades online.

Estudos nessa linha de pesquisa têm sido conduzidos em variados contextos: *newsgroup* (e.g. Usenet [37; 34; 9; 15; 38]), Wikipedia [36; 6; 21] e outros wikis (e.g. Cyclopath [30]), comunidades de desenvolvimento de *software open-source* [24], comunidades de compartilhamento de conteúdo [10; 5], sites de recomendação de filmes [11], e sites de Q&A (e.g. Yahoo! Answers (YA) [2], Naver Knowledge-iN [25] e StackOverflow [22]). Embora os perfis que melhor definem os contribuidores sejam contingentes do contexto de cada comunidade e do propósito dos analistas, algumas regularidades relevantes para este trabalho são

discutidas a seguir. Para evitar ambiguidades, referimos contribuidores que fornecem um grande volume de contribuições como ativistas, e contribuidores que geram contribuições de alta qualidade como *experts*.

Do ponto de vista da quantidade de atividade dos contribuidores, estudos em diversos sistemas têm geralmente observado comunidades formadas por uma ampla maioria de contribuidores pouco ativos e uma pequena fração de contribuidores ativistas [22; 25; 21; 38; 30; 10; 24]. No contexto de sites de Q&A, ativistas representam cerca de 1% dos usuários registrados, mas responsáveis por 22-28% das respostas desses sites [25; 22].

Uma segunda parte da literatura mostra que a distribuição das contribuições dos usuários ao longo do tempo também é frequentemente enviesada. Estudos nessa linha descrevem a atividade dos contribuidores com tipicamente um pico inicial de atividade seguido de uma queda acentuada [21; 30; 16]. Em sites de Q&A, Mamykina et al. [22] e Nam et al. [25] encontraram em diferentes comunidades um comportamento intermitente marcado por períodos de inatividade em responder. Além disso, Nam et al. observaram que o tempo que contribuidores estão ativos é positivamente correlacionado com a qualidade de suas respostas. A respeito da dinâmica do comportamento de grupos, Kittur et al. [21] observaram que grande parte das contribuições na Wikipedia e na rede social de favoritos Delicious, antes provida por ativistas, está gradualmente se modificando para um produto da atividade de usuários pouco ativos.

Diferenciando tipos de contribuição, é possível identificar grupos de usuários que são ativistas e/ou experts apenas em determinados tipos de contribuição. Identificar grupos de usuários nessa perspectiva pode, por sua vez, auxiliar o gerenciamento da comunidade e coordenar a atividade dos contribuidores. A Usenet foi amplamente estudado com diversos métodos e a literatura resultante consistentemente aponta para uma descrição dos participantes de *newsgroup* que incluem pessoas que apenas respondem/perguntam, *trolls* e *lurkers* [37; 34; 9; 15]. No contexto da Wikipedia, foi observado que contribuidores de diferentes perfis são responsáveis por produzir principalmente certos conteúdos [36] e que usuários inexperientes e experientes contribuem de maneira diferente [6]. Welser et al. [36] também mostram que nenhum desses perfis é formado em sua maioria por usuários experientes. Essa observação sugere que a Wikipedia não depende fortemente dos usuários experientes para o serviço

realizado por editores de nenhum dos perfis.

Especificamente analisando sites de Q&A, Nam et al. [25] sugerem que, assim como na Usenet, contribuidores podem ser claramente separados em *perguntadores* e *respondedores* no Naver Knowledge-iN. Adamic et al. [2] também identificaram esses dois perfis no Yahoo! Answers e ainda observaram categorias de questões com uma predominância de usuários *discutidores* – usuários que são ativos tanto em perguntar quanto em responder questões.

Outra tendência na análise de comportamento dos contribuidores é o estudo da estrutura da rede social criada, por exemplo, entre perguntadores e respondedores. Kang et al. [20] examinaram a influência da rede social de ativistas na qualidade de suas respostas. Usando dados do Yahoo! Answers e Naver Knowledge-iN, os autores descobriram que a quantidade de perguntadores ajudados influencia na qualidade de suas respostas, e que ativistas possuem uma relação de competição entre co-respondedores. Rodrigues et al. [32] estudaram sub-comunidades formadas nas categorias mais frequentemente utilizadas do Live QnA e do Yahoo! Answers, encontrando que ativistas podem estabelecer fortes laços sociais em determinadas categorias. Por fim, Adamic et al. [2] caracterizaram diversas categorias no Yahoo! Answers e identificaram que há uma diversidade de comportamentos quando comparados diferentes grupos de categorias (e.g. programação versus casamento) e que, ao lidar com questões mais objetivas, usuários mais focados tendem a receber mais votos positivos nas suas respostas.

## 2.2 Examinando ou identificando contribuidores experts

Uma vertente diferente de pesquisa foca em usuários de um perfil específico. Esse tipo de pesquisa geralmente visa examinar o comportamento de determinados usuários ou derivar heurísticas para predizer quais usuários atuarão de acordo com certo perfil. O perfil mais popular considerado nessa abordagem é o *ativista expert*, também conhecido por contribuidor "elite". Entender como tais usuários agem e automaticamente identificá-los podem novamente melhorar o escalonamento de tarefas ou de recomendação, e direcionar esforços na comunidade para promover a participação de contribuidores que produzem conteúdo de alta qualidade.

Estudos têm demonstrado que em alguns contextos, contribuidores elite possuem um

comportamento consistente desde o seu início no sistema [29; 30]. Além disso, diversos algoritmos têm sido desenvolvidos para identificar experts ou autoridades em certos temas na comunidade [27; 3].

Ao examinar sites de Q&A, estes trabalhos buscam predizer quais usuários são propensos a atuar como experts em certos temas. Por exemplo, Riahi et al. [31] e Hamrahan et al. [18] exploram meio automáticos para identificar os experts mais adequados para responder uma dada questão. Pal et al. [28; 26] mostraram que é possível predizer quais usuários se tornarão ativistas experts usando dados das primeiras semanas de atividade do usuário. Ao analisar a dinâmica do comportamento dos experts, Pal et al. [26] descrevem o comportamento desses usuários em três tipos: consistentemente ativo, inicialmente inativo e depois ativo, e o oposto.

## 2.3 Nossa contribuição

Este trabalho avança o atual conhecimento sobre comunidades de Q&A ao usar técnicas de análise multivariada para caracterizar os perfis comportamentais dos contribuidores. Cada perfil identificado representa uma co-ocorrência típica dos valores de um conjunto de métricas que descrevem a participação (quantidade) e habilidade (qualidade) dos usuários para múltiplos tipos de contribuição. Ao considerar um conjunto mais diverso de métricas, essa caracterização nos permite investigar, por exemplo, se contribuidores experts também são usuários ativos que criam um grande volume de respostas, questões ou comentários. Um retrato rico desses perfis possibilita o entendimento das diferentes habilidades dos contribuidores, dos objetivos e necessidades dos sites de Q&A (como também sugerido anteriormente por Gazan [14]). Essa caracterização informa o gerenciamento desses sites e o desenvolvimento de mecanismos para identificar experts e alocar tarefas na comunidade.

Nossa análise também contribui para ampliar o atual conhecimento sobre como o comportamento dos contribuidores muda ao longo do tempo. O presente entendimento de sites Q&A não descreve como o comportamento do usuário, segundo a quantidade e qualidade de diferentes tipos de contribuições, muda ao longo do tempo, ou como o site estruturalmente evolui com respeito aos perfis de seus contribuidores. Essa análise é necessária para identificar mudanças típicas de comportamento dos usuários e novamente tem implicações no gerenciamento dessas comunidades, apoiando o desenvolvimento de mecanismos para

promover ou inibir mudanças desejáveis ou indesejáveis de comportamento dos usuários.

Na Tabela 2.1 mostramos um comparativo dos trabalhos relacionados e posicionamos o nosso entre eles. Em trabalhos preliminares, avançamos esse conhecimento ao aplicar técnicas multivariadas para investigar o comportamento dos contribuidores em um [12] e em cinco sites de Q&A [13]. No presente estudo expandimos e refinamos essas análises. Primeiro, estendemos a generalização dos resultados utilizando novos dados da plataforma Stack Exchange que contemplam, além dos sites das análises anteriores, sites de diversos temas, tamanhos e idades. Segundo, revisamos o conjunto de métricas e as refinamos para obter um modelo mais simples e acurado dos perfis de contribuidores.

Tabela 2.1: Comparação dos trabalhos relacionados. Células marcadas com **X** indicam que o trabalho (linha) realizou a respectiva análise (coluna).

| Trabalho | Site | Dimensões do comportamento individual | | | Estrutura da comunidade |
|---|---|---|---|---|---|
| | | Quantidade | Qualidade | Dinâmica | |
| Mamykina et al. [22] | StackOverflow | **X** | - | **X** | **X** |
| Nam et al. [25] [1] | Naver KiN | **X** | - | **X** | **X** |
| Rodrigues et al. [32] | Live QnA | **X** | - | - | **X** |
| Adamic et al. [2] [2] | Yahoo! Answers | **X** | **X** | - | **X** |
| Kang et al. [20] | Naver KiN e YA | - | **X** | - | - |
| Riahi et al. [31] | StackOverflow | - | **X** | - | - |
| Hanrahan et al. [18] | StackOverflow | - | **X** | - | - |
| Pal et al. [28] [1] | TurboTax Live | **X** | **X** | - | - |
| Pal et al. [26] [1] | StackOverflow | **X** | **X** | **X** | - |
| Furtado e Andrade [12] | Super User | **X** | **X** | - | **X** |
| Furtado et al. [13] | 5 sites do SE | **X** | **X** | **X** | **X** |
| **O presente trabalho** | 36 sites do SE | **X** | **X** | **X** | **X** |

[1] Análise do comportamento dos usuários focada no grupo de contribuidores ativistas

[2] Análise da qualidade das contribuições focada nas respostas

# Capítulo 3

# Anatomia de sites de Q&A

Neste capítulo descrevemos com mais detalhes o modelo básico de funcionamento utilizado pela ampla maioria dos sites de Q&A. Em seguida, apresentamos a plataforma que hospeda os sites de Q&A estudados neste trabalho, o Stack Exchange (SE), juntamente da descrição do funcionamento e criação de seus sites.

## 3.1 Modelo básico de funcionamento

Sites de Q&A, tais como Yahoo! Answers, Quora, Naver Knowledge-iN, StackOverflow e demais sites da plataforma Stack Exchange, operam segundo um modelo comum de funcionamento. A Figura 3.1 mostra a página de uma questão com suas respostas e comentários na plataforma SE que segue esse modelo típico de site de Q&A. Utilizando essa página como exemplo, descrevemos a seguir o curso típico de eventos para que uma questão seja respondida nesse modelo de funcionamento:

1. Um usuário posta uma questão descrevendo um problema;
2. A pergunta criada é listada no site e visualizada por outros usuários, que podem votar sobre a utilidade ou não da questão, ou favoritá-la;
3. Um ou mais usuários postam respostas e comentários associados a questões e respostas, que também podem ser visualizadas e receber votos da comunidade. Comentários nesse contexto são utilizados como ferramenta para discussão e esclarecimento entre o usuário que postou a pergunta (perguntador) e a comunidade;

4. Por fim, o perguntador pode selecionar uma das respostas como a melhor.

Como resultado desse processo, cada questão, resposta e comentário possui um saldo de votos, baseados na quantidade de votos positivos e negativos recebidos, e no número de favoritos para questões. Os sites de Q&A geralmente usam esse saldo de votos das perguntas e respostas para definir a ordem em que as mesmas serão listadas no site.

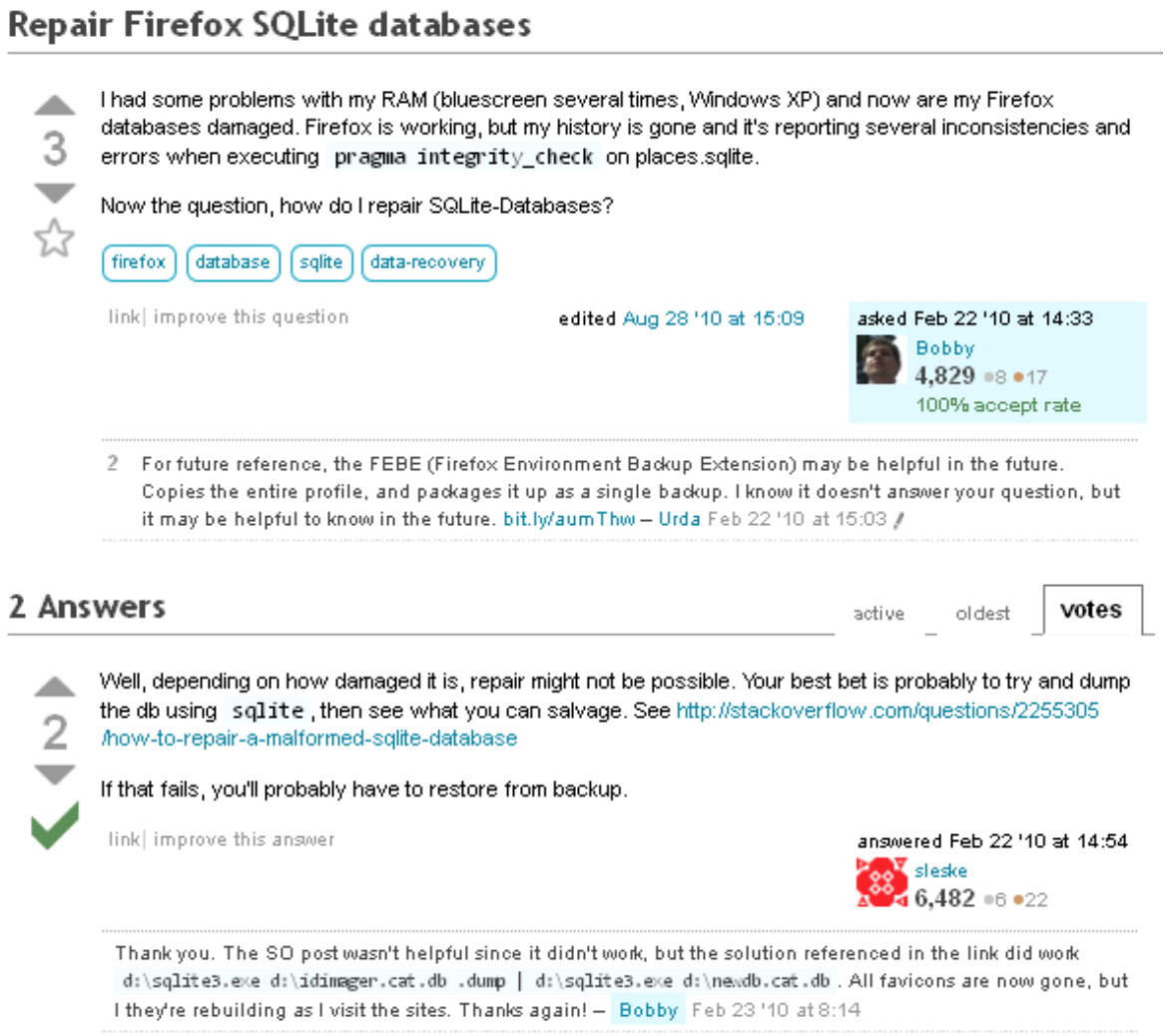


Figura 3.1: Página referente a uma questão da plataforma Stack Exchange.

## 3.2 A plataforma de Q&A Stack Exchange

A plataforma Stack Exchange é formada por uma rede de sites de Q&A onde cada site discute questões de um tópico específico. Criada a partir do sucesso do site original StackOverflow

(SO), seus criadores estenderam o modelo desse site com o objetivo de criar uma plataforma para a criação de novos sites. Atualmente a plataforma conta com 101 sites focados em questões que variam desde jogos e teoria da computação a culinária e fotografia. Com uma comunidade crescente, a plataforma agrega mais de 3 milhões de usuários registrados, mais de 11 milhões de respostas e recebe diariamente 8,4 milhões de visitas [8].

Novos sites no SE são criados numa parte da plataforma chamada de Area 51 [1]. Grupos de usuários se juntam nessa área da plataforma para propor e discutir ideias de novos sites. Ao atingir um número mínimo de usuários interessados numa dada proposta, o site proposto passa por uma fase beta. Nessa fase, o novo site é povoado com questões iniciais, um FAQ (dúvidas frequentes), moderadores temporários e a arte final de design. Passado o período probatório de 90 dias, se o site atingir algumas métricas de sucesso (e.g. um número mínimo de questões/visitas por dia e uma porcentagem de questões respondidas), ele é oficialmente lançado.

Os sites da plataforma operam de maneira independente e de forma similar ao modelo básico de funcionamento apresentado na Seção 3.1. Um novo usuário nessas comunidades pode desempenhar atividades como: postar perguntas e respostas, fazer edições e comentários nessas, definir qual das respostas é a melhor para suas perguntas e adicionar questões aos seus favoritos.

À medida que os usuários participam da comunidade e as suas contribuições são avaliadas positivamente, esses usuários são recompensados com pontos de reputação. Caso um usuário atinja uma quantidade preestabelecida de reputação, ele(a) ganha acesso a certas funcionalidades, que são geralmente meios para a realização de atividades de moderação.

Com base nessa descrição do funcionamento dos sites do SE, vemos que os usuários desses sites podem desempenhar diversas atividades. Esses usuários têm como atividades principais perguntar, responder, comentar e avaliar o conteúdo gerado. Realizando essas atividades, esses contribuidores colaboram para a criação de bases valiosas de conhecimento.

# Capítulo 4

# Perfis de contribuidores em 36 sites de Q&A

Nossas análises utilizam dados de 36 sites da plataforma de Q&A Stack Exchange. Neste capítulo apresentamos os sites e dados analisados, a abordagem para identificar os perfis de contribuidores e, por fim, os resultados e implicações dessa análise.

## 4.1 Dados analisados

A estratégia de nossa investigação é empírica e consiste em um estudo de caso que analisa dados de 36 sites da plataforma Stack Exchange. Na Tabela 4.1 apresentamos uma breve descrição desses sites. Apesar da plataforma hospedar atualmente 101 sites, seus administradores não disponibilizam os dados de sites menores em fase beta. Nossa análise não considera esses sites e nem o site mais popular da plataforma, o StackOverflow. Como o SO foi a instância original de onde o Stack Exchange foi generalizado, a sua história é mais longa e peculiar comparada aos sites restantes. Analisar tais mudanças no design da plataforma foge do escopo deste trabalho.

Tabela 4.1: Descrição dos 36 sites da plataforma Stack Exchange (dados de julho de 2012). Número de postagens é a soma das perguntas, respostas e comentários no site. Note que o número de contribuidores e postagens estão na escala de Milhar.

| Site | Tópico | # contrib. | # post. | Criação |
|---|---|---|---|---|
| Super User | Uso avançado de computadores | 66K | 748K | 2009-07 |
| Server Fault | Suporte e administração de sistemas | 54K | 707K | 2009-04 |
| Ask Ubuntu | Ubuntu | 36K | 298K | 2010-07 |
| Programmers | Desenvolvimento de software | 20K | 312K | 2010-09 |
| Mathematics | Matemática | 18K | 490K | 2010-07 |
| Meta | Melhorias para a plataforma SE | 14K | 387K | 2009-06 |
| Gaming | Jogos eletrônicos | 11K | 141K | 2010-07 |
| Apple | Uso de produtos da Apple | 11K | 106K | 2010-08 |
| English | Língua Inglesa | 9K | 185K | 2010-08 |
| Unix | Unix & Linux | 8K | 91K | 2010-08 |
| Wordpress | Plataforma Wordpress | 8K | 111K | 2010-08 |
| TeX | Sistema de tipografia | 7K | 179K | 2010-07 |
| WebMasters | Operação de Websites | 7K | 51K | 2010-07 |
| WebApps | Uso de aplicativos Web | 6K | 36K | 2010-06 |
| GameDev | Desenvolvimento de jogos | 6K | 78K | 2010-07 |
| Android | Uso de dispositivos Android | 5K | 44K | 2010-09 |
| Stats | Estatística | 5K | 85K | 2010-07 |
| Physics | Física | 4K | 88K | 2010-11 |
| UX | Interação Humano-Máquina | 4K | 50K | 2010-08 |
| Drupal | Plataforma Drupal | 4K | 67K | 2011-03 |
| SharePoint | Plataforma Microsoft SharePoint | 4K | 69K | 2009-10 |
| Electronics | Engenharia Elétrica e Eletrônica | 4K | 91K | 2009-10 |
| GIS | Sistemas de Informação Geográfica | 4K | 64K | 2010-07 |
| DBA | Uso avançado de bancos de dados | 4K | 47K | 2011-01 |
| Cooking | Cozinha | 4K | 51K | 2010-07 |
| Security | Segurança de TI | 3K | 39K | 2010-11 |
| Photo | Fotografia | 3K | 61K | 2010-07 |
| Sci-fi | Ficção científica | 3K | 49K | 2011-01 |
| DIY | Reformas domésticas | 3K | 34K | 2010-07 |

| | | | | |
|---|---|---|---|---|
| CS Theory | Teoria da computação | 2K | 34K | 2010-08 |
| Skeptics | Ceticismo científico | 2K | 35K | 2011-02 |
| Bicycles | Bicicletas | 1,6K | 23K | 2010-08 |
| RPG | Jogos de interpretação de personagens | 1,4K | 33K | 2010-08 |
| StackApps | Desenvolvimento usando a API do SE | 0,8K | 9K | 2010-05 |
| Judaism | Judaísmo | 0,7K | 52K | 2009-12 |
| Mathematica | Software Wolfram Mathematica | 0,7K | 28K | 2012-01 |

O Stack Exchange disponibiliza periodicamente bases de dados contendo a atividade – postagens de perguntas, respostas e comentários – de todos os usuários registrados desde o início de cada site[1]. Os dados históricos dos 36 sites compreendem desde o seu início até 31 de julho de 2012.

Para possibilitar a replicação do experimento, detalhamos no Apêndice D o ferramental utilizado e um passo-a-passo de como reproduzir esse experimento.

## 4.2 Métricas do comportamento do contribuidor

Considerando os dados apresentados dos 36 sites, nós extraímos os seguintes grupos de métricas para cada usuário, considerando um período de atividade:

**Métricas de participação**:

- *Número de perguntas* postadas em um período;
- *Número de respostas* postadas em um período;
- *Número de comentários* postados em um período; e
- *Tempo de atuação*, definido como o número de dias que o usuário realizou alguma atividade na comunidade – postando uma pergunta, resposta ou comentário.

[1] http://www.clearbits.net/creators/146-stack-exchange-data-dump

**Métricas de habilidade**:

- *Utilidade média das perguntas* (UMPerguntas): avalia o quanto a comunidade percebe a qualidade das questões do usuário. A qualidade de cada questão é medida como a soma do número de favoritos e seu saldo de votos. UMPerguntas de um usuário é a média da utilidade de todas as questões postadas por esse usuário.

- *Utilidade média das respostas* (UMRespostas): mede a habilidade do usuário em responder questões em comparação com as respostas de outros usuários que também responderam as mesmas questões. Para cada resposta que o usuário criou e para as quais existam respostas competidoras numa dada questão, a utilidade da resposta desse usuário é calculada por meio da normalização do saldo considerando todas as outras respostas competidoras nessa questão (i.e. calculamos seu z-score), ou é definida como zero, se nenhuma das respostas obteve votos. UMRespostas de um usuário é a média da utilidade das respostas postadas que têm respostas competidoras, ou zero se nenhuma das respostas do usuário teve competição. Para melhor avaliar a qualidade de uma resposta, adicionamos um voto positivo no cálculo do saldo de votos, se a resposta foi selecionada como a melhor.

- *Utilidade média dos comentários* (UMComentários): determina o quão útil a comunidade avalia os comentários do usuário. É calculado de maneira análoga a UMPerguntas mas considerando apenas os votos em comentários postados pelo usuário em questões e respostas de outros usuários. Os comentários que um usuário faz nas suas próprias questões são desconsiderados porque esses comentários geralmente não são avaliados pela comunidade. Observamos que esses comentários tipicamente estão relacionados à conversação, requisição ou esclarecimento, e não à informação.

## 4.3 Identificando perfis de longo prazo

Nossa primeira análise foca em explorar perfis de longo prazo típicos dos contribuidores dos 36 sites analisados. Esses perfis descrevem o comportamento dos contribuidores ao observar todo o período de atividade dos usuários nos sites estudados.

O problema de identificar perfis é equivalente ao de descobrir grupos de contribuidores com comportamento similar. Para abordar esse problema, empregamos técnicas de análise de agrupamento [4], que visam agrupar indivíduos maximizando, ao mesmo tempo, a homogeneidade interna dos grupos resultantes e a heterogeneidade entre esses grupos.

Nesta análise utilizamos o conjunto métricas de participação e habilidade que definimos e calculamos seu valor considerando o período que compreende toda a atividade de cada usuário. Contribuidores que não tiveram atividade antes do início do último mês da base de dados foram excluídos devido à pequena quantidade de informação disponível sobre seu comportamento.

Para definir o espaço de similaridade entre os usuários, utilizamos os valores normalizados (z-scores) das sete métricas que descrevem o comportamento do contribuidor. Dado que as escalas das métricas nas diferentes comunidades podem variar, a normalização é feita por site. Para comparar o comportamento de dois contribuidores utilizamos a medida mais comum de similaridade para análise de agrupamento [17] – a distância Euclidiana no espaço definido pelas métricas normalizadas.

## 4.4 Algoritmo de agrupamento

Dentre os algoritmos de agrupamento existentes na literatura, optamos por uma combinação dos algoritmos hierárquico e não-hierárquico para identificar os perfis de contribuidores nos dados. Por um lado, algoritmos hierárquicos têm a vantagem de serem independentes de parâmetros iniciais, o que permite o analista investigar uma variedade de soluções de agrupamento produzidas através da divisão e união iterativa ótima dos grupos. Algoritmos não-hierárquicos, por outro lado, são otimizados para encontrar a solução global, provendo soluções que são mais robustas a outliers que os métodos hierárquicos [23], mas que a qualidade depende de sementes iniciais dos centros dos grupos, e presume um número conhecido de grupos a serem descobertos.

Nossa análise combina essas duas abordagens usando primeiro o algoritmo de agrupamento de Ward [35] para explorar uma ampla variedade de soluções com diferentes quantidades de grupos. O resultado da exploração hierárquica nos informa um número adequado de grupos e seus centros, que são por sua vez usados como sementes dos centros no algoritmo

não-hierárquico k-means [19]. Lembramos que ambos os algoritmos de Ward e k-means são técnicas de agrupamento *hard*, portanto cada contribuidor está presente exatamente em um único grupo nos nossos resultados.

## 4.5 Definindo o número de grupos

O tamanho das 36 comunidades em conjunto – cerca de 359 mil contribuidores – torna impraticável rodar o algoritmo de Ward, uma vez que esse método exige a computação de uma matriz de similaridade NxN. Para abordar essa questão, executamos o algoritmo de Ward com amostras aleatórias dos contribuidores de cada um dos sites para estimar o número de grupos e seus centros, e em um segundo passo, usamos essa informação para executar um algoritmo mais escalável, o k-means, que possibilita o agrupamento de todos os usuários dos 36 sites.

Para prevenir um viés resultante do comportamento mais comum em usuários dos maiores sites, utilizamos amostras de tamanho igual para cada comunidade no algoritmo de Ward. O número de contribuidores da menor comunidade, Mathematica, limita o tamanho dessa amostra aleatória, o que resultou em uma amostragem total de 25.812 contribuidores extraídos de todos os sites.

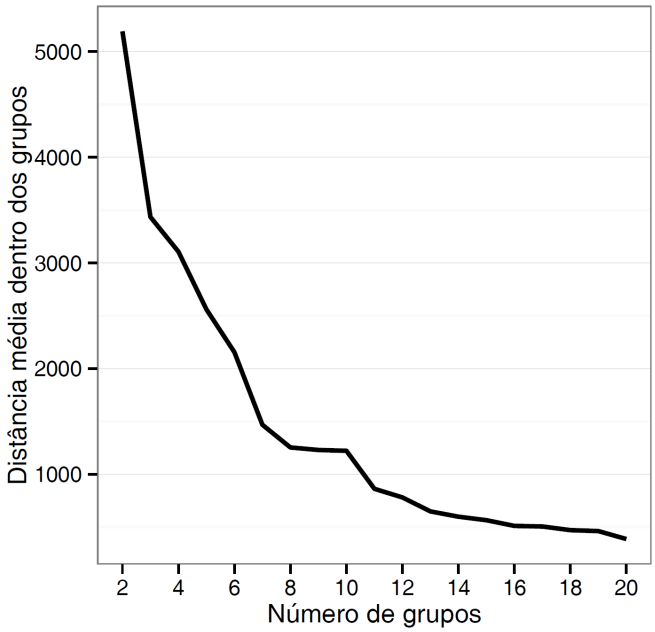


Figura 4.1: Análise da heterogeneidade obtida em cada solução de agrupamento.

A Figura 4.1 mostra a distância média dentro dos grupos para as diferentes soluções resultantes da execução do algoritmo de Ward nos nossos dados. É notável que a partir de 8 grupos as soluções começam a ganhar menos homogeneidade. Examinando mais detalhadamente as soluções com mais de 8 grupos, percebemos que a partir da solução com 10 grupos em diante, os novos grupos são notadamente similares aos encontrados previamente. Na Tabela 4.2 mostramos que os centros dos grupos que surgem nas soluções com 10 e 11 grupos são semelhantes aos centros 2 e 5 identificados na solução de 9 grupos. Escolhemos portanto a solução com 9 grupos como a mais representativa no nosso estudo porque cada grupo dessa solução possui comportamentos claramente distintos.

Tabela 4.2: Centros normalizados dos grupos identificados no agrupamento hierárquico. A primeira parte da tabela lista os centros da solução com 9 grupos, e a segunda lista os novos centros que surgem nas soluções de 10 e 11 grupos.

| Centros | Resp. | Perg. | Comen. | Tempo Atuação | UMResp. | UMPerg. | UMComen. |
|---|---|---|---|---|---|---|---|
| Centro 1 | -0,09 | -0,26 | -0,12 | -0,18 | -1,97 | -0,42 | -0,22 |
| **Centro 2*** | **-0,13** | **0,05** | **-0,09** | **-0,13** | **0,20** | **0,27** | **-0,19** |
| Centro 3 | -0,08 | -0,21 | -0,11 | -0,15 | -0,92 | -0,29 | -0,22 |
| Centro 4 | -0,13 | -0,22 | -0,12 | -0,21 | 0,22 | -0,50 | -0,24 |
| **Centro 5**** | **0,02** | **-0,19** | **-0,03** | **0,02** | **0,02** | **-0,13** | **1,59** |
| Centro 6 | -0,07 | -0,16 | -0,09 | -0,11 | 2,25 | -0,20 | -0,08 |
| Centro 7 | -0,09 | -0,04 | -0,07 | -0,09 | 0,12 | 3,81 | -0,07 |
| Centro 8 | 0,78 | 1,97 | 0,73 | 1,48 | 0,06 | 0,62 | 0,37 |
| Centro 9 | 5,81 | 4,80 | 5,48 | 6,77 | 0,41 | 0,86 | 0,59 |
| **Novo Centro - 10*** | **-0,15** | **-0,12** | **-0,12** | **-0,19** | **0,22** | **0,82** | **-0,22** |
| **Novo Centro - 11**** | **0,04** | **-0,19** | **-0,02** | **0,05** | **-0,03** | **-0,16** | **0,87** |

Informados por essa exploração, executamos o algoritmo k-means nos dados completos das 36 comunidades para identificar nove grupos, e fornecemos como semente inicial desse algoritmo os centros dos grupos identificados na análise do agrupamento hierárquico. Os grupos resultantes dessa análise são notavelmente semelhantes aos grupos encontrados no agrupamento hierárquico e os seus centros estão descritos na Figura 4.2. Na Tabela do Apêndice A.1 também expomos esses centros usando os valores não normalizados das mé-

tricas.

## 4.6 Rótulos dos perfis de contribuidores

Utilizando o centro dos grupos como sumário do comportamento do grupo, interpretamos nesta etapa da análise cada centro como um perfil de contribuidor no contexto de sites de Q&A. Os rótulos para os nove perfis encontrados e suas características marcantes estão descritos abaixo:

1. *Passageiro*: contribuidores com uma passagem breve na comunidade e que param de atuar sem demonstrar índices expressivos de participação e habilidade.

2. *Ocasional*: usuários que contribuem moderadamente, geralmente com questões, durante um tempo de atuação acima da média. Suas perguntas tendem a ser consideradas úteis.

3. *Imperito em respostas*: contribuidores com participação semelhante aos Passageiros (tempo de atuação abaixo da média), contudo as respostas desses contribuidores tendem a ser mal avaliadas pela comunidade.

4. *Expert em respostas*: usuários de atividade mediana, mas que possuem um excelente desempenho em contribuir com respostas de boa qualidade.

5. *Expert em perguntas*: contribuidores cujas perguntas são reconhecidas como importantes pela comunidade. Esses usuários são ligeiramente mais ativos que experts em respostas.

6. *Expert em comentários*: usuários com pouca atividade, mas que produzem comentários considerados úteis pela comunidade.

7. *Ativista em Q-A*: contribuidores que participam ativamente na comunidade, gerando uma quantidade significativa de respostas e principalmente perguntas, e que cuja habilidade em responder é ligeiramente abaixo da média.

8. *Ativista respondedor*: usuários com um longo tempo de atuação e um alto número de postagens, especialmente respostas. Além disso, as respostas desses usuários são

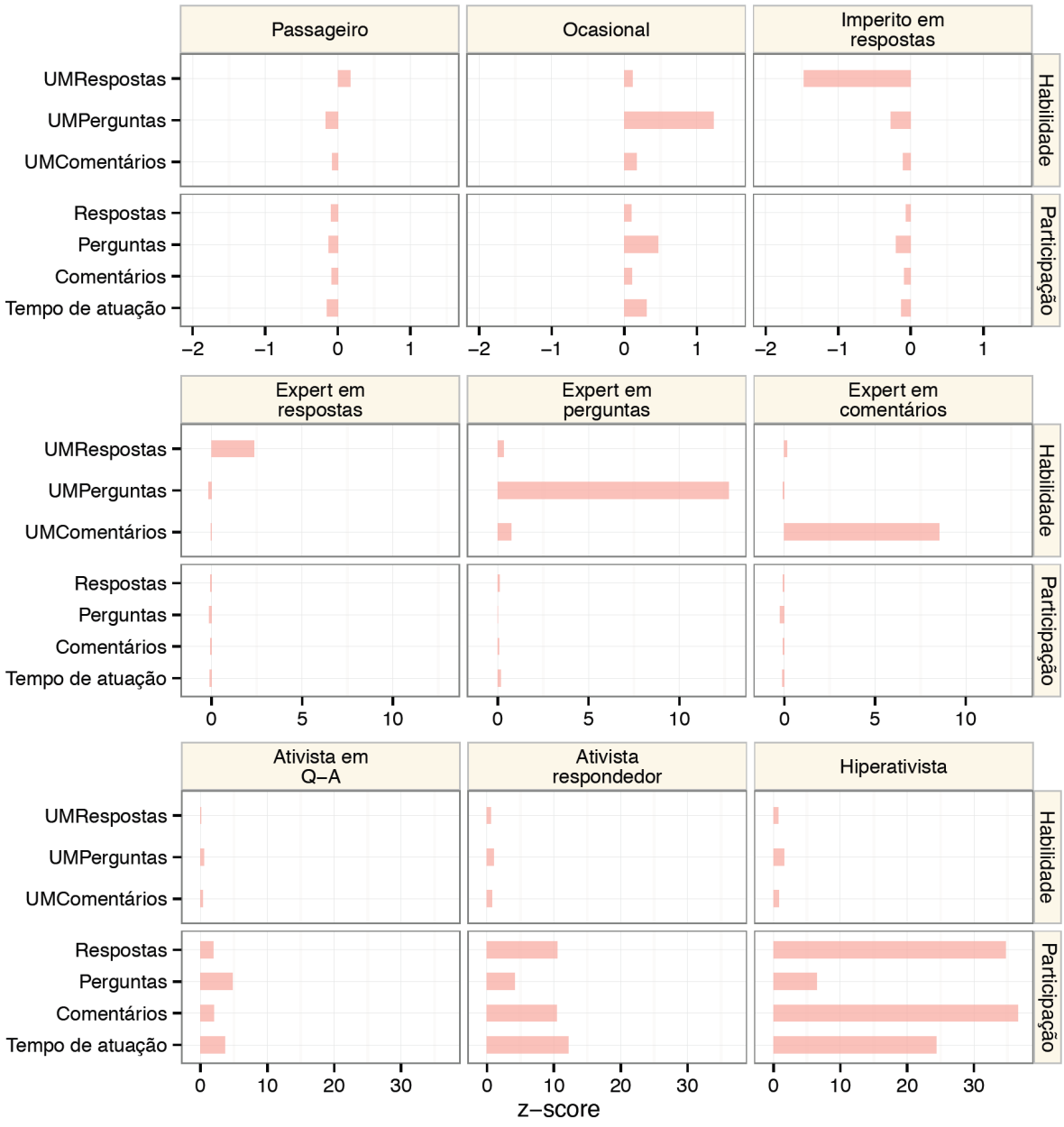


Figura 4.2: Centros dos nove grupos identificados. Note que o eixo horizontal é o z-score da métrica e as escalas nos três grupos de gráficos variam.

geralmente bem avaliadas. Os índices de participação indicam que esses usuários são altamente comprometidos com a comunidade, ao mesmo tempo que os índices de habilidade sugerem que as contribuições desses usuários são frequentemente úteis ao coletivo.

9. *Hiperativista*: contribuidores com um perfil semelhante aos ativistas respondedores, mas que contribuíram com um número desproporcional de respostas e comentários para o site. Entre os perfis identificados, esses usuários possuem o maior tempo de atividade nesses sites.

## 4.7 Discussão dos resultados

A análise de agrupamento utilizando dados de atividade de 36 sites de plataforma Stack Exchange revelou nove perfis de contribuidores. Esses perfis representam as combinações mais comuns de indicadores de participação e habilidade entre os usuários desses sites da plataforma. Focando nos perfis menos ativos identificados (todos exceto ativistas), observamos usuários que geram contribuições de qualidade média, experts em cada um dos tipos de contribuição, e contribuidores cujas respostas são mal avaliadas. Já contribuidores com perfil de alta a extrema atividade, por sua vez, têm habilidade menos acentuada do que seu nível de atividade. A Figura 4.3 sumariza essa visão.

A observação de que experts e contribuidores de alta atividade formam grupos disjuntos contribui para entender onde a *expertise* está localizada nesses sites, e tem implicações para mecanismos de alocação de tarefas. Primeiro, essa evidência sugere a necessidade de examinar se os atuais métodos desenvolvidos para identificar experts com alta atividade (e.g. [28; 26; 31; 18]) conseguem reconhecer com precisão os experts dos nossos resultados. Segundo, parece promissor usar mecanismos de alocação de tarefas para direcionar experts para responder principalmente perguntas difíceis quando eles atuam no sistema. Por fim, devido o seu baixo nível de atividade, mecanismos de alocação de tarefas deveriam considerar sugerir a mesma questão para múltiplos experts, ou para uma combinação de experts e de usuários de alta atividade para aumentar as chances de obter uma resposta em um tempo razoável.

O fato de que experts são tipicamente contribuidores menos ativos aponta para a necessidade de investigar suas motivações. Por um lado, se os gerentes desses sites promoverem

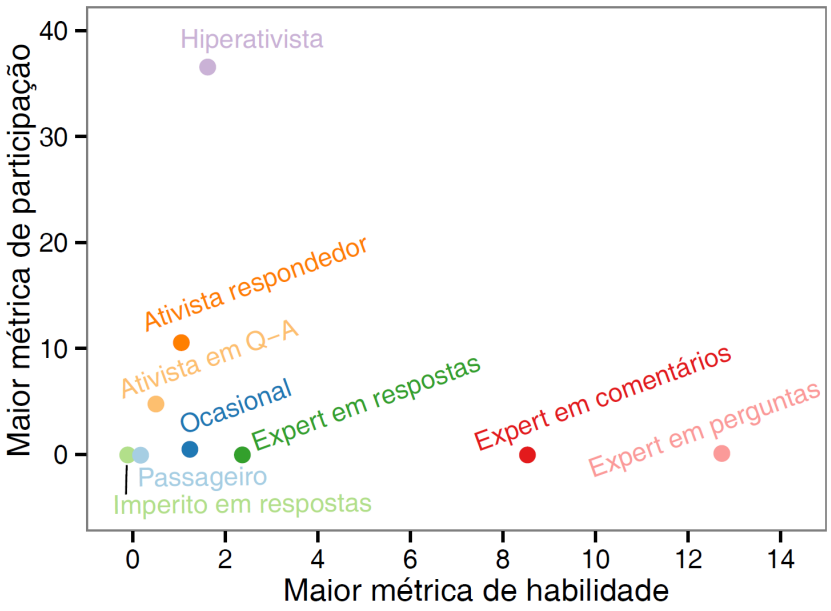


Figura 4.3: Gráfico de dispersão da maior métrica de participação e de habilidade dos centros dos grupos identificados.

a participação desses usuários, o serviço provido pela comunidade provavelmente terá um impacto positivo. Por outro lado, parece necessário entender até que ponto esses usuários são percebidos como experts porque eles são seletivos em responder questões. Não é óbvio que aumentar a atividade desses usuários irá necessariamente elevar proporcionalmente a qualidade das contribuições.

A proeminência do perfil imperito em respostas chama a atenção de projetistas e gerentes desses sites. Investigar os meios necessários para reduzir o efeito potencialmente negativo das contribuições de tais usuários pode melhorar o desempenho da comunidade como um todo. Além disso, a ausência de um grupo distinto de imperitos em respostas com alta atividade em nossa análise sugere que o design do Stack Exchange inibe a contribuição continuada de conteúdo percebido como fraco. Examinar quais mecanismos têm esse efeito pode ajudar a generalizar boas práticas para o planejamento de sites de Q&A.

Entre os perfis com alta atividade, percebemos que esses usuários se distinguem principalmente por perguntarem mais que a média (Ativistas de Q-A), por responder notavelmente mais que a média (Ativistas respondedores), e por um volume extremamente alto de contribuições em geral (Hiperativistas). Identificar tais usuários é relevante para alocar tarefas na comunidade, uma vez que esses são os contribuidores com a maior chance de respon-

der rapidamente a uma requisição ou sugestão. Além disso, hiperativistas não estão apenas contribuindo com conteúdo, mas também estão preocupados com o funcionamento da comunidade. Ao examinar a eleição de 2011 para moderadores da comunidade Super User, constatamos que um em seis hiperativistas se candidatou na eleição. Esse número é de uma ordem de magnitude maior que a de qualquer outro perfil se voluntariar na eleição. Identificar hiperativistas pode portanto ajudar os gerentes dessas comunidades a encontrar usuários dispostos a contribuir na moderação do site.

Nossos resultados também se assemelham em alguns aspectos com análises anteriores em sites de Q&A que examinaram perfis de acordo com o nível de atividade dos contribuidores [2; 25]. Os estudos prévios também identificaram perfis com tendências a perguntar e responder. Contudo, nossos resultados complementam essa caracterização ao considerar dimensões de qualidade e quantidade combinadas, revelando um conjunto de perfis mais diverso. Além disso, nossa análise revela o perfil imperito em respostas, que até então não tinha recebido muita atenção no contexto de sites de Q&A.

Comparando com nossos experimentos anteriores, percebemos que os nove perfis identificados são notavelmente semelhantes aos perfis observados no experimento em um [12] e em cinco sites de Q&A [13]. Encontrar perfis semelhantes em diferentes populações reforça a generalização dessa caracterização.

# Capítulo 5

# Composição dos sites

Neste capítulo utilizamos os perfis de contribuidores encontrados na análise de agrupamento para examinar a distribuição dos usuários nesses perfis nos 36 sites da plataforma Stack Exchange. Nosso objetivo é investigar a prevalência de contribuidores em cada perfil, e as semelhanças e diferenças nas composições das comunidades.

## 5.1 Descrição geral das composições

A Figura 5.1 mostra as distribuições dos perfis de contribuidores nos 36 sites. De forma geral, percebemos que quase a metade ou mais dos contribuidores desses sites (44-73%) se encaixam no perfil passageiro. O segundo e terceiro perfis mais frequentes nessas comunidades são os imperitos em respostas (13-26%) e contribuidores ocasionais (7-20%). Considerando esses 3 perfis, é possível notar que aproximadamente 90% dos contribuidores desses sites são formados por usuários de baixa a média atividade e por usuários com contribuições avaliadas como pouco úteis. Entre os contribuidores restantes, os perfis mais comuns são os experts em respostas (3-9%) e ativistas de Q-A (0,5-2,9%).

Por um lado, não é surpresa que experts e ativistas são menos comuns, dado que esses perfis exigem mais esforço e habilidade do usuário. Por outro lado, é surpreendente que imperitos em respostas sejam o segundo tipo mais comum de contribuidor nessas comunidades. Esse fato reforça a necessidade de investigar o efeito negativo que as contribuições de usuários imperitos podem causar no serviço desses sites.

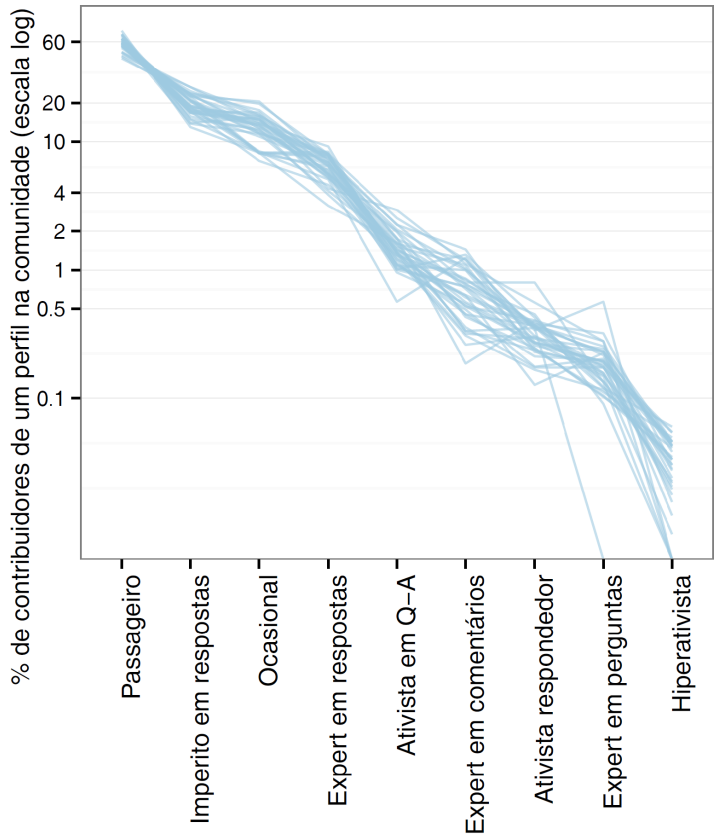


Figura 5.1: Distribuições dos perfis de contribuidores nos 36 sites do Stack Exchange. Cada linha representa a composição de um site.

## 5.2 Análise das variações de composição

Para verificar a variação das composições dos sites, testamos a independência entre a ocorrência de um perfil e o site onde o usuário participa através do teste Chi-quadrado de Pearson. Visando atender a premissa do teste Chi-quadrado de um número mínimo de observações nas classes da variável perfil, foi necessário excluir dessa análise usuários experts em perguntas, dada sua rara ocorrência, e unir em uma única classe hiperativistas e ativistas respondedores. Ainda com esse tratamento, sites menores – Mathematica, Judaism, StackApps, RPG e Bicycles – continuaram sem atender o número mínimo de observações e tiveram que ser removidos dessa análise. Considerando os demais sites, o resultado do teste Chi-quadrado aponta uma associação significativa entre os perfis e os sites ($\chi^2(180) = 7556,07, p <,001$), indicando que apesar de uma notável semelhança na ordem em que os perfis ocorrem, as comunidades apresentam variações significativas de composição.

Aprofundando na análise das variações de composição, interpretamos os resíduos obtidos

do teste Chi-quadrado[1]. A análise dos resíduos mostra que sites onde usuários passageiros são mais frequentes – Mathematics, Meta, Stats, Ask Ubuntu e TeX – tendem a ter menos contribuidores imperitos, experts em respostas e, em parte desses sites, menos ocasionais. Uma explicação é que os tópicos discutidos nesses sites limitaram a participação da massa de usuários menos motivada – contribuidores de baixa a média atividade.

Para ajudar a interpretação desse resultado, a Figura 5.2 mostra a dispersão do percentual desses perfis em função de dois tipos de tópico, tema casual (conhecimento comum) e tema especializado (restritos a uma área de conhecimento). Comunidades como Mathematics, Stats e TeX discutem tópicos mais restritos ao meio acadêmico e Meta, site de propostas de melhorias para a plataforma, é de forte interesse para usuários mais engajados. A menor atividade da massa de usuários pouco ativos intuitivamente explicaria a menor frequência de imperitos, experts em respostas e ocasionais, uma vez que os mesmos são perfis com a atividade de responder ligeiramente mais alta que passageiros.

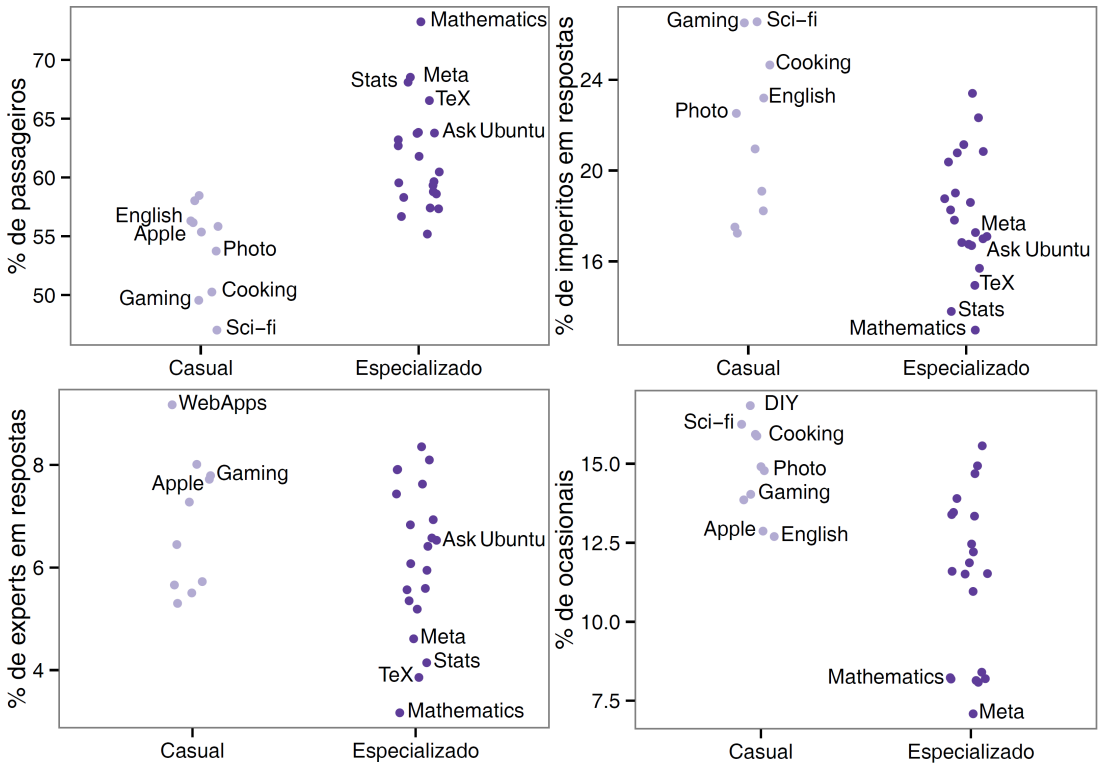


Figura 5.2: Gráfico de dispersão do percentual de passageiros, ocasionais, imperitos e experts em respostas, em função do tipo de tópico discutido na comunidade.

[1]Para fins de clareza na explicação, comentamos apenas os resultados significativos dessa análise, e apresentamos todos os resíduos numa tabela no Apêndice B.1.

No outro sentido, temos comunidades com menos passageiros – Sci-fi, Gaming, Cooking, Apple, English e Photo – e, em contrapartida, mais ocasionais, imperitos e, em parte desses sites, experts em respostas. Contrapondo o padrão anterior, a discussão de temas mais casuais, e em certos casos de domínio mais amplo, pode ter facilitado uma maior atividade da massa de usuários menos ativa.

A observação de que as frequências de passageiros, ocasionais, imperitos e experts se relacionam com os temas da comunidade tem implicações para o gerenciamento desses sites. Se por um lado as comunidades de tópicos mais casuais se beneficiam de uma maior participação da massa de usuários menos ativa, revelando inclusive em parte desses sites contribuidores com alta habilidade em prover respostas, por outro lado aparentemente também contribui para que exista nessa massa um maior contingente de contribuidores com atividade marcada por uma predominância de respostas inadequadas.

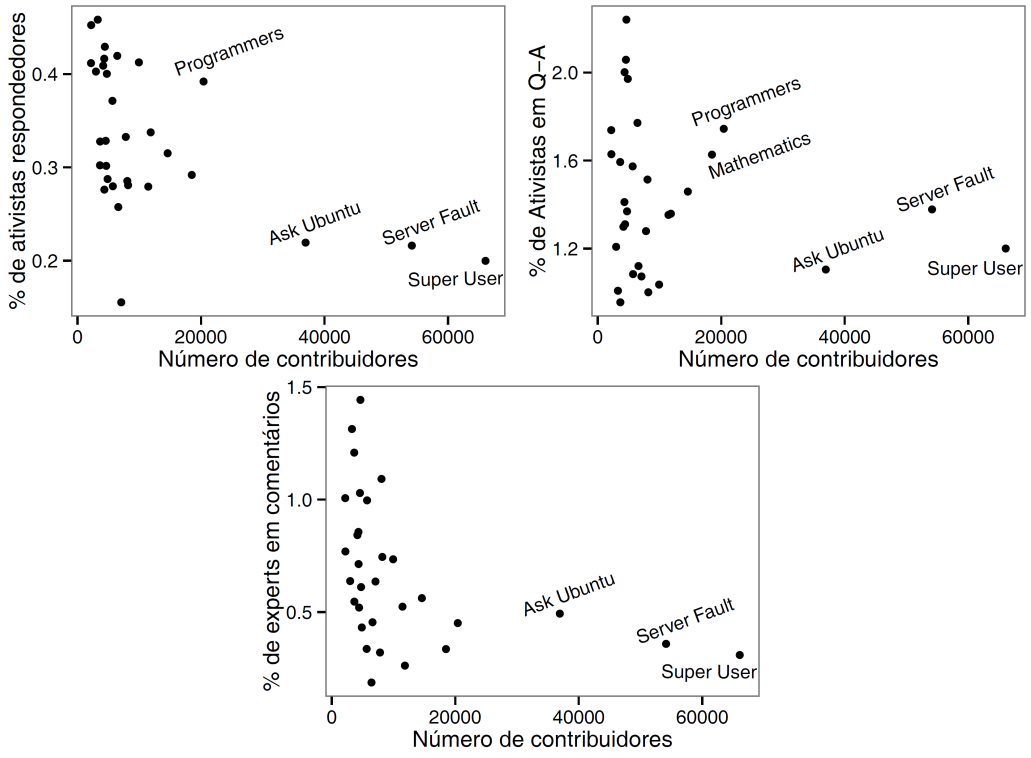


Figura 5.3: Gráfico de dispersão do percentual de ativistas em Q-A, respondedores (mais hiperativistas) e experts em comentários, em função do tamanho da comunidade.

A análise dos resíduos de perfis menos comuns juntamente com o gráfico de dispersão da Figura 5.3 mostram uma tendência dos maiores sites de nossa análise – Super User, Server Fault e Ask Ubuntu – a apresentar uma menor proporção de ativistas respondedores, discu-

tidores (Q-A) e experts em respostas. Esse resultado reforça a rara ocorrência desses perfis e sugere que a partir de certa escala, a velocidade do crescimento desses grupos é menor em comparação aos demais perfis.

Analisando os resíduos também identificamos sites que se destacam por uma alta frequência de ativistas discutidores em sua composição, a exemplo das comunidades Mathematics, Game Dev, Programmers, GIS, Physics, SharePoint e Drupal. Em relação a ativistas respondedores, apenas três sites apresentam com uma ocorrência significativamente mais alta desses usuários – Game Dev, Programmers e English.

Curioso que Programmers e Mathematics estão entre os maiores sites de nossa análise, mas ainda assim engajam uma quantidade relativamente alta de ativistas. Os dados que usamos não explicam esse fenômeno nesses sites, contudo esse fato deve motivar trabalhos futuros para investigar as motivações de ativistas nessas comunidades. Esse estudo pode revelar lições importantes para os outros sites da plataforma.

## 5.3 Discussão dos resultados

De forma geral, percebemos que embora a ordem de ocorrência dos perfis nos 36 sites seja notavelmente semelhante, as composições dos sites variam de forma significativa. Em nossa análise mostramos como aspectos do tópico e tamanho dos sites se correlacionam com o tipo de usuário que os mesmos atraem com o tempo. Nessa análise observamos composições de comunidade marcadas por (I) um grande volume de passageiros, (II) um número reduzido de passageiros e em compensação um elevado número de imperitos em respostas, e (III) pela relativa baixa ocorrência de ativistas.

Gerentes de comunidades de Q&A devem tomar cuidado ao lidar com sites de tópicos casuais porque, embora a massa de usuários tenda a ser ligeiramente mais participativa, esses sites tendem a atrair usuários com contribuições mal avaliadas. Ao lidar com sites de maior porte, gerentes devem se atentar para uma redução no crescimento do grupo de ativistas.

Trabalhos futuros nessa linha deveriam considerar investigar porque ativistas nos sites Programmers e Mathematics fogem deste padrão de redução na proporção de ativistas. Esse estudo pode revelar um conhecimento útil ao desenvolvimento de mecanismos para motivar mais ativistas nos demais sites da plataforma.

# Capítulo 6

# Produção dos perfis

Nos capítulos anteriores discutimos sobre os perfis de comportamento dos contribuidores e a composição das comunidades segundo esses perfis. Todavia, essas análises não esclarecem o papel dos grupos formados por esses perfis na construção da base de conhecimento dos sites. Por exemplo, a importância dos grupos formados por perfis de baixa (e.g. passageiros e ocasionais) e de alta atividade (e.g. ativistas) na criação de perguntas, respostas e comentários.

Neste capítulo esclarecemos essa questão examinando a produção dos grupos[1] nos 36 sites. Para realizar essa análise, definimos a produção dos grupos em termos da quantidade e qualidade das contribuições agregadas ao sistema. Para quantidade, consideramos o volume de perguntas, respostas e comentários produzido pelo conjunto de contribuidores de cada perfil. Para qualidade, consideramos o número de votos positivos recebidos nas perguntas, respostas e comentários. Para facilitar a leitura dos resultados, nos referimos a usuários passageiros e ocasionais, em conjunto, como contribuidores sem habilidade marcante.

## 6.1 Descrição geral das produções

A Figura 6.1 apresenta uma visão geral da produção dos grupos nos 36 sites. Em geral, percebemos que ambos imperitos e experts produzem em conjunto um volume pequeno das contribuições, e recebem uma fração pequena das avaliações positivas – imperitos recebem menos votos, e experts mais votos, em comparação ao volume de contribuições do grupo.

[1] O termo grupo é usado no restante deste trabalho para referir a um conjunto de usuários de mesmo perfil.

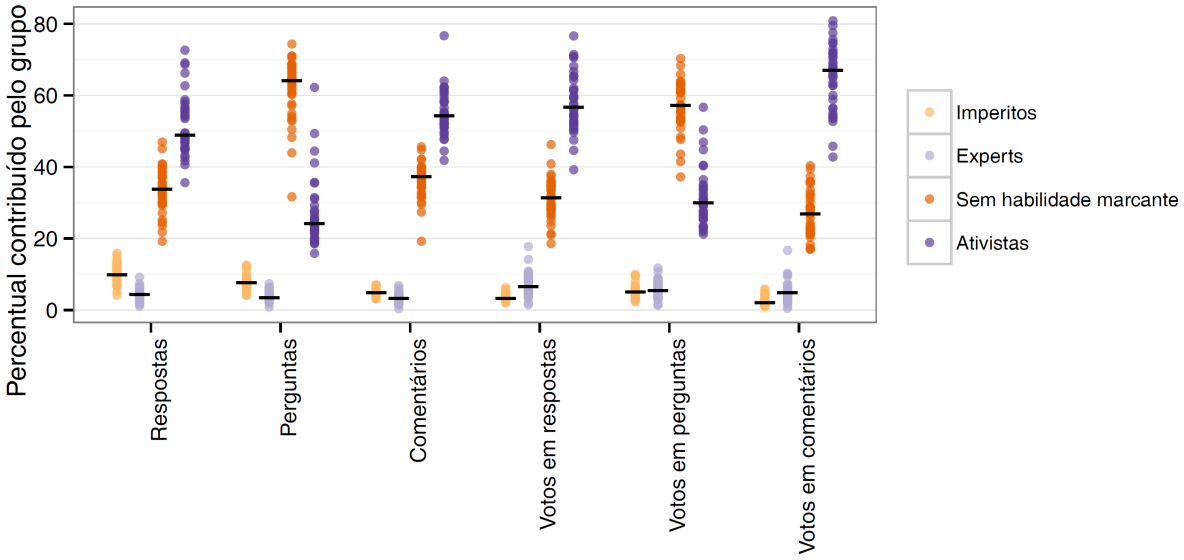


Figura 6.1: Parcela de contribuição agregada dos grupos formados pelos diferentes perfis nos 36 sites. Note que as barras pretas são as medianas dos percentuais observados nesses sites. Os perfis foram agrupados para facilitar a legibilidade e o rótulo do perfil *sem habilidade marcante* se refere aos contribuidores passageiros e ocasionais.

Contribuidores sem habilidade marcante e ativistas, por sua vez, criam a maioria do conteúdo nesses sites. Ativistas são responsáveis por grande parte das respostas e comentários, e recebem também a maior parte dos votos nesses dois tipos de contribuição. Por outro lado, contribuidores sem habilidade marcante contribuem com a maioria das perguntas e recebem grande parte dos votos nessas.

Essa descrição geral das produções nos 36 sites revela uma separação clara entre o grupo que desempenha fundamentalmente o papel de responder/comentar e o grupo que cria a maioria das perguntas. Interessante que contribuidores sem habilidade marcante além de desempenharem um papel mais proeminente na criação de perguntas, também contribuem junto aos ativistas na geração de uma parcela significativa de respostas e comentários.

## 6.2 Análise das variações de produção

Examinando a produção dos perfis em cada site individualmente, identificamos comunidades com características peculiares de produção. Essa análise revela quatro tipos de produção

notadamente distintos da mediana dos sites. A Figura 6.2 mostra o resultado de quatro sites[1] dos conjuntos de sites formados pelos tipos de produção identificados. Características e exemplos de outros sites que apresentam esses tipos de produção são descritos a seguir.

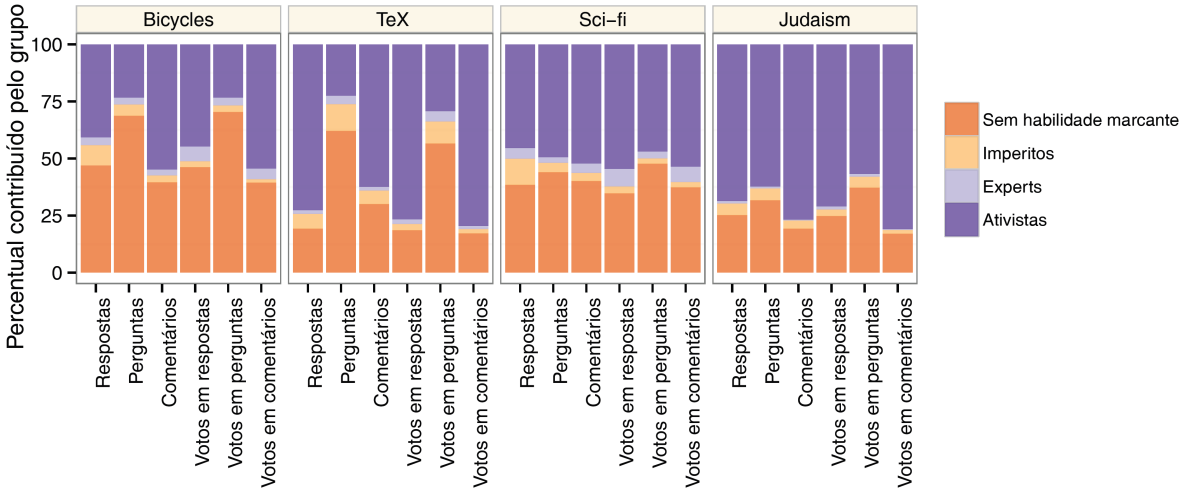


Figura 6.2: Produção dos grupos em quatro sites com características de produção diferentes da mediana dos sites.

Em relação à produção mediana, as comunidades Bicycles, RPG, Programmers, Computer Science Theory, WebApps e Cooking se destacam por uma participação ligeiramente mais alta, em relação aos demais sites, de contribuidores sem habilidade marcante na criação de respostas. O fato do nascimento do Stack Exchange ter base no StackOverflow pode explicar os sites Programmers e Computer Science Theory estarem nesse grupo. É possível que seja comum os usuários da plataforma dominarem assuntos teóricos relacionados à computação. Já em relação aos sites Bicycles, RPG, WebApps e Cooking, podemos especular que o tema mais casual pode ter facilitado a atividade desses usuários. A comunidade WebApps, por exemplo, discute o uso de aplicativos Web (i.e. Gmail e Facebook), tópico que certamente a maioria dos contribuidores desses sites domina.

De modo contrário ao padrão anterior, o volume de respostas, comentários e votos recebidos é fortemente concentrado em ativistas nas comunidades TeX, Meta, Mathematics, StackApps e Electronics. A separação clara de usuários que só respondem ou só perguntam nesses sites pode ser um efeito do tema dominado por um grupo restrito de usuários. A comunidade Meta, por exemplo, é dedicada a usuários com o interesse de propor melhorias

[1] Os resultados para todos os sites são detalhados no Apêndice B.3.

para a plataforma e como visto na discussão dos perfis no Capítulo 4, o pequeno grupo de hiperativistas é fortemente interessado no funcionamento dos sites.

Em contraste com a separação dos papéis dos grupos, ativistas nas comunidades Sci-fi, Gaming, RPG e Skeptics juntamente com contribuidores sem habilidade marcante desempenham ao mesmo tempo os papéis de gerar e responder perguntas. É possível inferir que tópicos menos objetivos, mais abertos à discussão, tenham estimulado a geração de perguntas e respostas em proporções semelhantes.

Por último, identificamos um site que combina dois padrões apresentados anteriormente. Ativistas e contribuidores sem habilidade marcante da comunidade Judaism desempenham papéis similares em perguntar e responder, e ao mesmo tempo, ativistas desproporcionalmente dominam a produção de respostas, perguntas e comentários, e de votos recebidos. A comunidade pequena, o tema de interesse restrito e o caráter religioso são fatores que podem ter estimulado alta participação de ativistas nesse site.

## 6.3 Discussão dos resultados

Nossa análise revela que os diferentes papéis assumidos por usuários ativistas e contribuidores sem habilidade marcante podem ser decorridos do tipo de tópico discutido na comunidade. Adamic et al. [2] também identificaram variações semelhantes nos papéis assumidos por usuários nas categorias do Yahoo! Answers. Categorias relacionadas a assuntos mais subjetivos, tais como religião, política e esportes, tendem a se parecer com fóruns de discussão, com uma alta proporção de usuários perguntando e respondendo ao mesmo tempo. Por outro lado, categorias que discutem tópicos com respostas objetivas – programação, ciência e matemática – são marcadas por uma maior separação dos papéis das pessoas que só perguntam ou só respondem. O primeiro tipo se assemelha as comunidades Judaism, Sci-fi, Gaming, RPG e Skeptics, enquanto que o segundo se assemelha a sites como TeX, Meta, Mathematics, Electronics e, em menor intensidade, aos demais.

Considerando as regularidades observadas, nossa análise revela que o conteúdo nesses 36 sites é produzido em sua ampla maioria por dois grupos: uma grande base de contribuidores que atuam esporadicamente e não tem uma habilidade marcante, e um núcleo menor e mais ativo de contribuidores. No cenário comum, o primeiro tem um impacto significa-

tivo na geração de perguntas, enquanto o segundo tem uma notável importância em ajudar a comunidade com respostas e comentários de qualidade. É importante notar também que contribuidores sem habilidade marcante ainda proveem uma fração considerável das respostas e comentários e recebem uma parcela significativa dos votos nessas contribuições.

Experts e imperitos são de limitada importância dado o número escasso de contribuições e de votos recebidos. No entanto, apesar de pequena, a fração de respostas mal avaliadas criadas por imperitos ($\sim$10%) não pode ser considerada desprezível.

Ainda sobre as regularidades, nossos resultados para produção de respostas dos perfis são similares aos observados por Mamykina et al. [22]. Eles também observaram que contribuidores de alta e baixa atividade têm importâncias semelhantes na produção de respostas no site StackOverflow. Nossos resultados complementam essa análise ao mostrar como esse padrão ocorre para questões e comentários, e para aspectos da qualidade das contribuições. Além disso, mostra sites com configurações de produção particulares.

De forma geral, nossa caracterização indica que gerentes de sites de Q&A devem dar atenção não somente a ativistas, mas também a usuários passageiros e ocasionais para manter esses sites produtivos. Na maioria dos sites, usuários de alta atividade juntamente com contribuidores sem habilidade marcante são importantes na produção de perguntas, respostas e comentários.

Nossos resultados ainda apontam que o volume de conteúdo criado por usuários imperitos é limitado, mas não desprezível. Mecanismos que ajudem a controlar e/ou melhorar a qualidade de suas contribuições pode ter um impacto positivo nesses sites.

Considerando sites com produções dos grupos diferentes do retrato geral, nossos resultados dão indícios de como o tema da comunidade influencia a produção dos grupos, mostrando que em sites de temas de conhecimento mais geral há uma maior participação de contribuidores de atividade esporádica, e que em sites de conhecimento mais restrito as contribuições são concentradas em ativistas. Essa análise também mostra que a subjetividade do tema está relacionada com contribuidores de alta e de esporádica atividade produzindo conteúdo em proporções semelhantes.

Esses resultados indicam a necessidade de entender os riscos e benefícios dos diferentes tipos de produção para a operação desses sites. Nossa análise deixa espaço para investigar as vantagens e desvantagens para o funcionamento de um site de Q&A quando a comuni-

dade apresenta uma separação clara dos papéis dos grupos em gerar e responder perguntas, e quando esses papéis são mais igualmente distribuídos. Além disso, cenários em que o site depende fortemente de um pequeno núcleo de usuários ativistas para a produção de conteúdo, assim como a comunidade Judaism, podem apresentar riscos à manutenção do serviço oferecido pelo site.

# Capítulo 7

# Dinâmica dos perfis de contribuidores

As análises apresentadas nos capítulos anteriores descrevem perfis de comportamento que consideram toda a atividade dos contribuidores. Embora essas análises apresentem um resumo relevante do comportamento do usuário, elas não informam sobre como o comportamento desses usuários muda ao longo do tempo. Neste capítulo investigamos o aspecto temporal do comportamento dos usuários caracterizando perfis de curto prazo – perfis que descrevem a atividade dos usuários em janelas curtas de tempo.

Nosso objetivo é examinar a dinâmica de comportamento dos contribuidores e a evolução de propriedades estruturais do site, como sua composição. Para realizar essa análise, conduzimos uma versão longitudinal da nossa caracterização no maior dos 36 sites estudados, Super User. Consideramos apenas esse site nessa análise devido à complexidade para realizar um estudo longitudinal.

## 7.1 Identificando perfis de curto prazo

Os dados históricos coletados do Super User compreendem 37 meses de atividade do site. Desse total decidimos excluir da análise os últimos sete meses de atividade do site para evitar um viés na identificação de abandono do usuário do sistema, dado que o mesmo pode deixar de contribuir nos últimos meses dos dados, mas voltar a contribuir nos meses seguintes. Os dados desse período – que chamamos de período de confirmação de inatividade – são utilizados somente para checar se o usuário ainda estará ativo depois de certo ponto da análise.

Usando os 30 meses iniciais de atividade do site, discretizamos os dados em 15 janelas de 2 meses. A escolha desse tamanho de janela se deve ao tempo médio de permanência dos usuários no site que é de quatro meses e meio, e ao tempo mediano de nove meses e meio. Ao considerar janelas de 2 meses, podemos observar o comportamento da maioria dos contribuidores em mais de uma unidade de análise.

Um contribuidor é considerado ativo em uma janela de tempo se ele(a) contribuiu com pelo menos uma questão, pergunta ou comentário naquela janela. Contribuidores cuja atividade se inicia perto do fim de uma janela – após o fim do primeiro mês – são excluídos da análise da mesma para prevenir a subestimação de sua atividade, devido ao curto período de observação. Esses contribuidores, no entanto, poderão ser considerados ativos nas janelas seguintes independentemente da localização de sua atividade na janela. Realizando a análise dessa forma, é possível que observemos um contribuidor iniciando sua atividade no final de uma janela e só voltando a atuar no final de outra. Contudo, apesar de suas contribuições se encontrarem no fim dessa última janela, esse usuário teve um período superior a um mês para demonstrar claramente algum comportamento na comunidade.

Assim como feito na análise de perfis usando todo o período de atividade, para cada janela de tempo calculamos a normalização z-score das métricas de participação e habilidade dos contribuidores considerando a atividade dos usuários ativos naquela janela. Após esse pré-processamento, conduzimos a análise de agrupamento dos contribuidores ativos usando a mesma combinação de métodos hierárquicos e não-hierárquicos aplicados na análise dos dados completos dos 36 sites.

O volume de atividade dos usuários nas 15 janelas – cerca de 72 mil observações – também nos impossibilitou de executar o algoritmo hierárquico de Ward com todos os dados. Similarmente a análise anterior, resolvemos esse problema colhendo amostras aleatórias de mesmo tamanho para cada uma das janelas. A amostragem de 2 mil usuários de cada janela resultou numa amostra de 30 mil observações (41,6% do total). Após estimar os centros e a quantidade de grupos através do algoritmo de Ward, executamos o algoritmo não-hierárquico k-means usando os dados completos.

Os resultados dessa análise de agrupamento são notavelmente similares aos da análise considerando todo o período de atividade dos sites. Os mesmos nove perfis que caracterizaram o comportamento de longo prazo também conseguem descrever a atividade de curto

prazo dos contribuidores (o Apêndice C detalha esses resultados). Também é válido notar que o fato de que identificamos os mesmos perfis considerando diferentes períodos de tempo e em diferentes populações sugere a generalidade dos resultados dessa caracterização.

## 7.2 Evolução da composição do Super User

A análise de agrupamento considerando janelas de 2 meses nos informa como a distribuição dos perfis de contribuidores evolui ao longo do tempo no Super User. A Figura 7.1 ilustra essa evolução mostrando a proporção de contribuidores de cada perfil em cada janela de tempo. Como se pode notar a proporção de usuários na grande maioria dos perfis é estável ao longo do tempo. As únicas exceções a essa estabilidade são as proporções de usuários passageiros e experts em respostas. Contribuidores passageiros aumentam em cerca de 10% no final do segundo e início do terceiro ano (12º e 13º janelas) dos nossos dados, enquanto que experts em respostas consistentemente decrescem ao longo desse período.

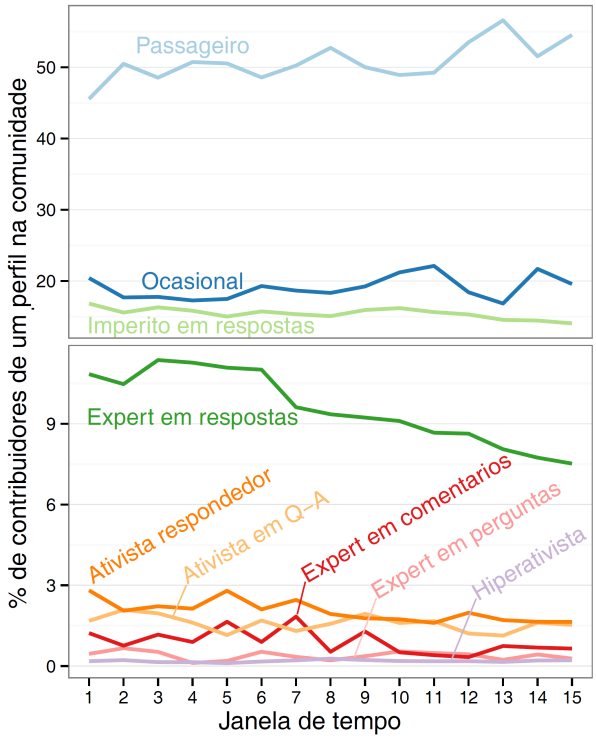


Figura 7.1: Composição da população de contribuidores ativos do Super User ao longo do tempo. Note que as escalas são diferentes nas duas partes do eixo vertical.

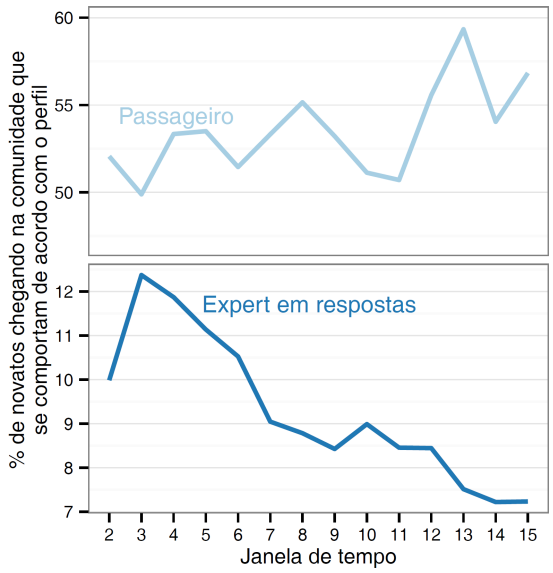


Figura 7.2: Percentual de novatos chegando na comunidade que se comportam de acordo com os perfis cujas proporções mudam significativamente ao longo do tempo na comunidade. Note que as escalas são diferentes nas duas partes do gráfico.

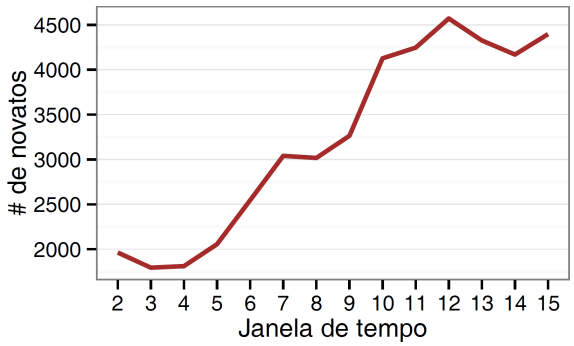


Figura 7.3: Quantidade de novatos se juntando a comunidade Super User ao longo do tempo.

Aprofundando nessa análise, observamos que essas tendências estão intimamente relacionadas ao tipo de novato que o site atrai com o tempo. A Figura 7.2 mostra que a probabilidade de um novo contribuidor de um determinado perfil se juntar à comunidade em diferentes períodos acompanha as mudanças na composição da Figura 7.1. Ao longo do tempo, contribuidores novatos tendem a atuar em menor probabilidade como experts em respostas. Já a tendência crescente do perfil passageiro entre novatos se inicia próximo do fim do segundo ano (12º janela). Essa tendência também se inicia no mesmo período em que a comunidade começa a receber um volume grande de novos contribuidores (Figura 7.3).

A observação que a chegada de um número elevado de novos contribuidores na comuni-

dade é correlacionada com a mudança nos perfis mais frequentemente assumidos por novatos é de interesse dos gerentes dessas comunidades e merece investigações adicionais. É possível que o Super User tenha começado a atrair um público diferente com o passar do tempo. Contudo, também é possível que o alto número de contribuidores chegando tenha dificultado a geração de respostas de alta qualidade por novatos, implicando na diminuição na proporção de experts em respostas na comunidade. Relacionado a esse resultado, trabalhos na Wikipedia têm reportado um aumento constante no número de contribuidores de baixa atividade ao longo do tempo [21].

A relativa estabilidade observada na evolução da composição do Super User também contribui para facilitar o nosso estudo de dinâmica de mudança de perfil dos usuários. Nessa análise, apresentada a seguir, não foi necessário isolar períodos específicos de anomalia para estudar o comportamento dos contribuidores na comunidade.

## 7.3 Dinâmica de mudança de perfil

Após examinar a dinâmica da composição do Super User, voltamos nossa análise para investigar a dinâmica dos perfis de contribuidores. Nosso primeiro experimento nessa análise objetiva medir a probabilidade de mudança entre dois perfis quaisquer durante janelas consecutivas de tempo. Nesse processo, também calculamos a probabilidade de um contribuidor se tornar inativo na próxima janela de tempo, ou abandonar o site. Consideramos que um contribuidor abandonou o site se esse contribuidor não realizar contribuições em janelas futuras, incluindo o período de confirmação de inatividade. Um contribuidor é considerado inativo se o mesmo não criou respostas, perguntas ou comentários numa dada janela de tempo e ainda não abandonou o site.

A probabilidade de mudança entre perfis, e de cada perfil para um estado inativo ou de abandono da comunidade são apresentadas na Figura 7.4. Esse resultado mostra que contribuidores de todos os perfis, exceto ativistas, tendem a reduzir sua atividade nas janelas seguintes de tempo. Grande parte desses usuários se tornam inativos, passageiros ou contribuidores ocasionais. Diferentemente, ativistas são perfis mais estáveis, e frequentemente mantêm seu comportamento entre janelas consecutivas de tempo. Essa observação aponta que a proporção de contribuidores de alta atividade tende a ser constante ao longo do tempo

e que há uma limitada renovação desse grupo de usuários.

| Perfil | H | AR | AQA | ER | EP | EC | IR | O | P | INA | AC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Hiperativista (H) | .56 | .35 | .00 | .02 | .02 | .00 | .01 | .02 | .02 | .02 | .00 |
| Ativista respondedor (AR) | .03 | .46 | .01 | .08 | .00 | .01 | .07 | .07 | .20 | .04 | .02 |
| Ativista em Q-A (AQA) | .00 | .03 | .31 | .03 | .00 | .00 | .04 | .34 | .15 | .07 | .03 |
| Expert em respostas (ER) | .00 | .01 | .00 | .06 | .00 | .01 | .06 | .06 | .19 | .29 | .32 |
| Expert em perguntas (EP) | .00 | .03 | .01 | .03 | .00 | .00 | .05 | .12 | .25 | .28 | .24 |
| Expert em comentários (EC) | .00 | .01 | .00 | .06 | .00 | .02 | .06 | .05 | .23 | .35 | .20 |
| Imperito em respostas (IR) | .00 | .00 | .00 | .03 | .00 | .00 | .06 | .05 | .15 | .25 | .45 |
| Ocasional (O) | .00 | .01 | .02 | .03 | .00 | .00 | .04 | .15 | .21 | .28 | .25 |
| Passageiro (P) | .00 | .00 | .00 | .03 | .00 | .00 | .04 | .06 | .17 | .30 | .39 |
| Inativo (INA) | .00 | .00 | .00 | .03 | .00 | .00 | .03 | .05 | .17 | .71 | .00 |

Figura 7.4: Probabilidade de mudar de um perfil X (linha) para um perfil Y (coluna). Portanto, a soma de cada linha é 1. A coluna AC contém as probabilidades de abandonar a comunidade. Usuários que não atuaram numa dada janela são considerados Inativos.

Os resultados da nossa análise ainda reforçam estudos anteriores da dinâmica de atividade em sites de Q&A. Esses estudos têm consistentemente reportado que contribuidores de alta atividade tendem a ter um comportamento constante, enquanto que a maioria dos contribuidores apresentam picos de atividade seguidos por períodos de inatividade [22; 25; 26].

Focando nos perfis de alta atividade, vemos que hiperativistas e ativistas respondedores são intimamente relacionados. Hiperativistas tendem a se tornar ativistas respondedores, enquanto que esse último é o único perfil com chance de seus usuários se tornarem hiperativistas. A dinâmica de ativistas em Q-A, por outro lado, é notavelmente diferente. Esses usuários têm uma pequena chance de mudar para o perfil ativista respondedor e hiperativista, e apresentam uma probabilidade significativamente alta de se tornar contribuidores ocasionais. Essa observação indica que ativistas cujo foco não é responder e comentar tendem a demonstrar uma atividade mais localizada no tempo.

Considerando novamente o retrato geral, vemos que a maioria dos contribuidores além de frequentemente se tornarem inativos numa janela de tempo, também tendem a permanecer nesse estado por períodos consecutivos. Entre os usuários que voltam a atuar na comunidade, grande parte atua em perfis de baixa a média atividade nas janelas seguintes de tempo. Atuando em perfis baixa atividade, tais como passageiro, contribuidores em geral tendem a

abandonar a comunidade nas janelas seguintes.

Interessante que contribuidores experts mostram baixa estabilidade. Experts em respostas, em especial, apresentam uma alta probabilidade de abandonar o site. Similarmente, contribuidores imperitos em respostas também não tendem a permanecer contribuindo no mesmo perfil ao longo do tempo e exibem a maior probabilidade entre os perfis de abandonar o site. Quando não deixam o site, esses usuários se tornam tipicamente contribuidores passageiros ou ocasionais.

Nossos resultados em conjunto apontam que além de experts serem mais comuns entre usuários de pouca atividade, a atuação desses usuários tende a ocorrer em períodos localizados. Experts em respostas especialmente tendem a atuar por um curto período e a abandonar o site. Esse resultado indica a necessidade de investigar como aumentar a fidelidade de tais contribuidores. Além disso, mecanismos para identificar experts devem ser capazes de identificá-los durante as suas primeiras atividades no site, visto que esses usuários abandonam o site em um curto período de tempo.

A observação que imperitos em respostas têm a maior chance de abandonar a comunidade sugere que prover contribuições mal avaliadas é desmotivante para uma parcela significativa de contribuidores. Por um lado, esse fenômeno pode limitar a quantidade de conteúdo produzido que foge dos padrões da comunidade. Por outro, a alta proporção de imperitos presente na comunidade (Figura 7.1) indica que o Super User pode se beneficiar da aplicação de melhores mecanismos para educar os usuários. Prevenir a criação de respostas mal avaliadas pode aumentar a retenção dos usuários dispostos a contribuir para o site.

## 7.4 Perfil de novatos vs. experientes

Nosso último experimento examina como a probabilidade de atuação nos perfis varia em função da experiência do contribuidor no site. Para analisar o comportamento de novatos, examinamos o primeiro perfil assumido por todos os contribuidores. Para contribuidores experientes, analisamos o quinto perfil assumido por contribuidores que foram ativos em pelo menos cinco janelas de tempo. Usando esses dados aplicamos o teste Chi-quadrado de Pearson para verificar se há uma associação entre o perfil assumido pelo usuário e o fato dele ser novato ou experiente. A Tabela 7.1 mostra os resultados dessa comparação. Note

que devido ao pequeno número de hiperativistas, combinamos novamente os usuários desse grupo com ativistas respondedores para atender às premissas do teste Chi-quadrado.

Tabela 7.1: Resultado do teste Chi-quadrado de Pearson para verificar a associação entre a experiência (novato ou experiente) e o perfil assumido pelo contribuidor. Hiperativistas e ativistas respondedores são analisados em conjunto para atender as premissas do teste. Células em negrito indicam que a proporção encontrada na amostra foi significativamente diferente da proporção esperada caso não houvesse associação entre as variáveis ($p <, 001$).

| Chi-quadrado | | H + AR | AQA | ER | EP | EC | IR | O | P |
|---|---|---|---|---|---|---|---|---|---|
| Novato | Percentual observado | **1.11** | **0.90** | 8.88 | 0.33 | 0.48 | **18.76** | 17.87 | 51.66 |
| | Resíduo padrão | **-3.52** | **-2.53** | -1.23 | 0.17 | -1.79 | **3.01** | -1.58 | 0.78 |
| Experiente | Percentual observado | **3.41** | **2.38** | **11.00** | 0.28 | **1.23** | **11.48** | **21.72** | **48.47** |
| | Resíduo padrão | **10.84** | **7.80** | **3.80** | -0.51 | **5.52** | **-9.26** | **4.87** | **-2.39** |

O teste Chi-quadrado aponta uma associação significativa entre o fato do usuário ser novato ou experiente e o seu perfil assumido na comunidade ($\chi^2(7) = 374, 63, p <, 001$). Os resíduos desse teste mostram que usuários experientes são mais propensos a atuar em perfis mais ativos em comparação aos novatos. Contribuidores experientes apresentam uma alta chance de atuar como ativistas respondedores/hiperativistas e ativistas em Q-A, e são mais propensos a assumir o perfil ocasional e menos propensos a atuarem como passageiros.

Analisando os perfis experts, vemos que experientes apresentam uma probabilidade significativamente alta de atuar como experts em respostas e comentários. Enquanto isso, novatos não apresentam tendências significativas para atuar nos perfis de alta habilidade e são mais propensos a se comportar como imperitos em respostas do que experientes.

Nossos resultados têm semelhanças com as observações feitas por Welser et al. [36] em seu estudo na Wikipedia. Contudo, enquanto Welser et al. observaram que a Wikipedia não parece depender de usuários experientes para qualquer papel na comunidade, vemos que, no Super User, perfis de alta atividade e habilidade são fortemente relacionados com a experiência do usuário. Similarmente ao nosso estudo, Nam et al. [25] observaram uma correlação positiva entre a quantidade de períodos ativos do contribuidor e a qualidade de suas respostas.

Por fim, o fato de novatos não apresentarem uma menor tendência em atuar como experts,

mas serem mais propensos a se comportar como imperitos em respostas, é de interesse de gerentes de sites de Q&A. Esse resultado sugere que orientar novatos em suas primeiras respostas pode aumentar sua retenção na comunidade e diminuir o volume de conteúdo mal avaliado no site.

# Capítulo 8

# Conclusão

Neste capítulo apresentamos inicialmente um resumo da nossa caracterização do comportamento de contribuidores em sites de Q&A. Baseados nessa caracterização, discutimos como os resultados obtidos podem ajudar a melhorar o design de sites de Q&A e, por fim, apontamos limitações e direções em que nossa análise pode ser expandida.

## 8.1 Resumo

Neste trabalho apresentamos uma caracterização do comportamento dos contribuidores em sites de Q&A que descreve a quantidade (indicador de participação) e a qualidade (indicador de habilidade) de suas contribuições, e o aspecto da dinâmica de comportamento dos usuários. Na primeira etapa dessa análise, identificamos perfis comportamentais de longo prazo realizando uma análise de agrupamento dos dados de todo o período de atividade dos usuários de 36 sites da plataforma Stack Exchange. Usando esses perfis, mostramos o aspecto geral da composição e produção dos grupos nesses sites, e analisamos sites que apresentaram variações significativas nessas análises.

Na segunda etapa deste estudo, investigamos o comportamento dinâmico do contribuidor no maior dos 36 sites analisados, Super User. Para realizar essa análise, repetimos a análise de agrupamento usando dados de atividade dos usuários em 15 janelas de 2 meses para identificar perfis de curto prazo. Baseado na classificação dos perfis nessas janelas, observamos como a composição do Super User evolui ao longo do tempo e examinamos como o comportamento dos contribuidores muda entre janelas consecutivas. Por fim, nosso último

experimento verifica a relação entre a experiência do usuário no site e o perfil assumido por ele.

## 8.2 Resultados principais e implicações

Ambas as análises, de perfis de longo e de curto prazo, revelaram nove perfis que ajudam a melhorar a compreensão geral de como sites de Q&A funcionam. Examinando esses perfis, identificamos contribuidores de quatro tipos gerais: (I) contribuidores *sem habilidade marcante*, passageiro e ocasional; (II) *imperitos* em respostas; (III) *experts*, em respostas, em perguntas e em comentários; e (IV) *ativistas*, em Q-A, respondedor e hiperativista.

De maneira geral, conhecer esses perfis pode auxiliar o desenvolvimento de estratégias de gerenciamento nesses sites. Particularmente, observar que experts e contribuidores de alta atividade formam grupos distintos é um conhecimento útil para o desenvolvimento de mecanismos de alocação de tarefas e para a identificação de experts.

A análise de composição e produção dos grupos nos 36 sites em conjunto mostram que esses sites, em geral, são mantidos na sua maioria por uma massa de contribuidores sem habilidade marcante e que atuam esporadicamente, e por um pequeno grupo de contribuidores de alta atividade. Esse retrato dos sites aponta a importância não somente de ativistas, mas de contribuidores comuns sem habilidade marcante para gerar conteúdo para os sites.

O aspecto geral da nossa análise de produção e composição também chama atenção para experts e imperitos em respostas. A relativa pequena importância de experts na criação de conteúdo pode motivar gerentes de sites de Q&A a incentivar a participação desses usuários. O aumento da participação de experts pode elevar a qualidade geral das contribuições do site. Quanto a imperitos em respostas, parece promissor prover uma orientação adequada a esses usuários na criação de suas respostas. Imperito em respostas é o segundo tipo de contribuidor mais comum nesses sites, e a sua parcela de produção de respostas não é desprezível, portanto orientar esse tipo de usuário de forma a melhorar a qualidade de suas respostas pode melhorar significativamente o conteúdo produzido nesses sites.

Analisando composições com características particulares, observamos que variações de composição do site se relacionam com fatores como tamanho e o tipo de tópico discutido no site. Ao tratar de comunidades de tópicos casuais, gerentes de sites de Q&A devem tomar

cuidado com a participação ligeiramente mais alta da massa de usuários de pouca atividade. Apesar desse efeito contribuir em revelar, em parte desses sites, mais contribuidores de alta habilidade, esse efeito também contribui em atrair usuários com contribuições mal avaliadas. Esse resultado indica que mecanismos de orientação e redução de respostas de baixa qualidade são particularmente relevantes para sites de tópicos mais casuais.

Em relação ao fator tamanho, identificamos que sites de maior porte tendem a apresentar uma menor ocorrência de perfis menos comuns, como experts em comentários e ativistas. Dada a importância de ativistas, gerentes devem atentar para um possível impacto que a redução do percentual de usuários ativistas pode causar no funcionamento do site.

Quanto às variações de produção dos grupos, observamos que a produção de certos perfis se relaciona com o tipo de tópico discutido. Sites de tópicos casuais ou de conhecimento restrito implicam numa maior ou menor (resp.) participação de contribuidores sem habilidade marcante na produção de conteúdo nesses sites. Enquanto isso, a abertura à discussão do tópico, se esse discute questões objetivas ou subjetivas, impacta na maior ou menor (resp.) separação dos papéis das pessoas que só perguntam ou só respondem. Sites que discutem tópicos mais subjetivos envolvendo muita opinião, como religião e ficção científica, tendem a ter ativistas e contribuidores sem habilidade marcante contribuindo coletivamente com proporções semelhantes de perguntas, respostas e comentários. Consideradas essas variações, nossa análise abre espaço para investigar possíveis efeitos que diferentes tipos de produção dos grupos podem causar na operação dos sites.

Na perspectiva de dinâmica, a análise da evolução da composição do Super User mostra uma notável estabilidade da proporção dos perfis ao longo do tempo. Contudo, identificamos uma tendência no aumento de contribuidores passageiros e um decrescimento dos experts em respostas. Nossa análise indica que essa tendência está relacionada ao tipo de novato que o site atrai com o tempo. Essa evidência pode estar relacionada com uma mudança no público do site e carece de uma investigação adicional.

Examinando a dinâmica de comportamento dos usuários, identificamos que o pequeno grupo de ativistas apresenta comportamento estável, enquanto que os usuários dos demais perfis são de comportamento mais volátil ao longo do tempo. Esse resultado complementa múltiplos estudos que reportaram que a atividade da maioria dos contribuidores acontece em períodos localizados e indica a necessidade de investigar como manter esses usuários

contribuindo por mais tempo nesses sites.

Importante também notar que imperitos em respostas e experts tendem a abandonar a comunidade cedo. Esse resultado reforça a necessidade de investigar como aumentar a retenção de usuários que atuam como experts, e como melhor integrar contribuidores imperitos.

Finalmente, nossa comparação das distribuições dos perfis nas populações de novatos e experientes revela que o Super User parece depender de usuários experientes para desempenhar os perfis ativistas, expert em respostas e em comentários. Novatos, por sua vez, não apresentam uma significativa menor chance de atuar como experts, mas tendem a atuar como imperitos em respostas. Essa observação reforça novamente que fornecer auxílio para os usuários, principalmente novatos, em suas primeiras respostas pode contribuir para a redução de conteúdo mal avaliado no site.

## 8.3 Ameaças à validade

No nosso trabalho identificamos decisões feitas no *design* do experimento que podem ameaçar à validade dos resultados. A seguir, listamos essas decisões e as relacionamos aos tipos de ameaças à validade definidos por Cook & Campbell [7].

**Validade de conclusão**. Verificamos com o teste Chi-quadrado de Pearson que certos perfis tendem a ocorrer em maior e menor frequência em determinados sites. Contudo, não sabemos se há uma correlação estatística entre as composições e produções dos perfis com o tamanho e tipo de tópico do site.

**Validade interna**. As ameaças à validade interna pela seleção dos usuários é mínima, uma vez que todos os contribuidores dos sites foram incluídos no estudo. Nossos dados contemplam também toda a atividade do usuário em perguntar, responder e comentar desde a sua chegada na comunidade até a data da geração da base de dados.

**Validade de construção**. A medição da habilidade dos contribuidores por meio dos votos recebidos pode não estimar com precisão essa característica do comportamento dos usuários. Além disso, outros métodos de agrupamento utilizando outras métricas de similaridade, além da distância Euclidiana, podem revelar perfis diferentes de usuários. Trabalhos futuros deveriam considerar um design de experimento que compara os resultados obtidos variando as métricas de comportamento, a métrica de similaridade e o método de agrupamento.

**Validade externa**. Apesar dos sites do SE operarem segundo um modelo comum de funcionamento de sites de Q&A, o tamanho de suas comunidades é de porte inferior a outros sites – por exemplo, em comparação a sites mais populares como o Yahoo! Answers. Além disso, sites fora da plataforma SE possuem políticas próprias de moderação do conteúdo que podem afetar os comportamentos típicos dos usuários nessas comunidades.

## 8.4 Trabalhos Futuros

Extensões do nosso trabalho devem focar em investigar as razões por trás dos diversos comportamentos dos contribuidores. Nossa análise quantitativa evidencia a necessidade de investigar a motivação de usuários que se encaixam nos perfis experts e imperito em respostas. Essa investigação em conjunto com um estudo qualitativo de todos os perfis identificados pode contribuir para criar um retrato mais claro e rico dos contribuidores de sites de Q&A.

Analisando a composição e a produção dos grupos, observamos variações nesses resultados que podem motivar análises futuras para investigar os possíveis efeitos dessas variações no funcionamento dos sites. Um estudo poderia correlacionar sites com diferenças significativas de composição com métricas de performance dos sites. Na mesma linha, um novo estudo poderia examinar a correlação entre sites com diferentes tipos de produção dos grupos e métricas de desempenho desses sites.

Ainda na análise de composição, identificamos sites que apesar do relativo grande porte motiva uma quantidade significativamente alta de ativistas. Entender porque as comunidades Programmers e Mathematics apresentam essa característica pode ajudar a incentivar mais contribuidores de alta atividade em outros sites.

Relacionado à análise dinâmica, novos estudos deveriam considerar em expandir a generalização dos nossos resultados analisando sites de diferentes temas na plataforma Stack Exchange. Essa generalização também pode ser aprimorada com novas experimentações da dinâmica de comportamento considerando janelas de granularidade menor que 2 meses.

Novas análises de dinâmica também podem investigar como possíveis mudanças na composição e na produção dos grupos ao longo do tempo afetam o desempenho do site. Essa análise pode informar gerentes sobre os riscos desses eventos ao funcionamento do site.

Por fim, outra direção em que nosso estudo pode ser expandido seria estender toda a

caracterização dos perfis para sites de diferentes plataformas. Comparar o resultado dessa nova análise com o do presente estudo pode revelar informações importantes sobre o funcionamento de sites de Q&A e novas lições de como melhorá-los.

# Bibliografia


[1] Area 51. Stack exchange network staging zone. http://area51.stackexchange.com/, Maio 2013.

[2] Lada A Adamic, Jun Zhang, Eytan Bakshy, and Mark S Ackerman. Knowledge sharing and yahoo answers: everyone knows something. *Proceeding of the 17th international conference on World Wide Web*, Beijing, C:665–674, 2008.

[3] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 207–218, New York, NY, USA, 2008. ACM.

[4] Mark S Aldenderfer and Roger K Blashfield. *Cluster Analysis*, volume 07-044 of *Sage University Paper Series on Quantitative Applications in the Social Sciences*. Sage, 1984.

[5] Nazareno Andrade, Elizeu Santos-Neto, Francisco Brasileiro, and Matei Ripeanu. Resource demand and supply in BitTorrent content-sharing communities. *Comput. Netw.*, 53(4):515–527, March 2009.

[6] Susan L Bryant, Andrea Forte, and Amy Bruckman. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Human Factors*, volume 6 of *Net communities*, pages 1–10. ACM, ACM, 2005.

[7] T.D. Cook and D.T. Campbell. *Quasi Experimentation: Design & Analysis Issues for Field Settings*. Hougton Mifflin, 1979.

[8] Stack Exchange. Stack exchange about. http://stackexchange.com/about, Maio 2013.

[9] D Fisher, M Smith, and H T Welser. You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences HICSS06*, 00(C):59b–59b, 2006.

[10] F Font, G Roma, P Herrera, and X Serra. Characterization of the Freesound Online Community. In *Third International Workshop on Cognitive Information Processing*, Baiona, Spain, 2012.

[11] Paul Fugelstad, Patrick Dwyer, Jennifer Filson Moses, John Kim, Cleila Anna Mannino, Loren Terveen, and Mark Snyder. What Makes Users Rate ( Share , Tag , Edit ...)? Predicting Patterns of Participation in Online Communities. In *Communities*, CSCW '12, pages 969–978. ACM, 2012.

[12] Adabriand Furtado and Nazareno Andrade. Ativistas, passageiros, ocasionais e especialistas: Perfis de usuário na construção de um site de Q&A. In *Anais do VIII Simpósio Brasileiro de Sistemas Colaborativos*, 2011.

[13] Adabriand Furtado, Nazareno Andrade, Nigini Oliveira, and Francisco Brasileiro. Contributor profiles, their dynamics, and their importance in five Q&A sites. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, pages 1237–1252, New York, NY, USA, 2013. ACM.

[14] Rich Gazan. Social Q&A. *Journal of the American Society for Information Science and Technology*, 62(12):2301–2312, 2011.

[15] Scott A Golder and Judith Donath. Social roles in electronic communities. *Internet Research*, 5:1–25, 2004.

[16] Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang, and Yihong (Eric) Zhao. Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 369–378, New York, NY, USA, 2009. ACM.

[17] Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson. *Multivariate data analysis (7th Edition)*. Prentice Hall Higher Education, 2010.

[18] Benjamin V Hanrahan, Gregorio Convertino, and Les Nelson. Modeling problem difficulty and expertise in stackoverflow. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, CSCW '12, pages 91–94, New York, NY, USA, 2012. ACM.

[19] J A Hartigan and M A Wong. A K-Means Clustering Algorithm. *Applied Statistics*, 28(1):100–108, 1979.

[20] Minhyung Kang and Byoungsoo Kim. Understanding the Effect of Social Networks on User Behaviors in Community-Driven Knowledge Services. *Journal of the American Society for Information Science and Technology*, 62(6):1066–1074, 2011.

[21] A Kittur, E Chi, B A Pendleton, B Suh, and T Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *Algorithmica*, 1(2):1–9, 2007.

[22] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. Design Lessons from the Fastest Q & A Site in the West. In *Human Factors*, CHI '11, pages 2857–2866. ACM Press, 2011.

[23] Glenn W Milligan. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3):325–342, 1980.

[24] Kumiyo Nakakoji, Yasuhiro Yamamoto, Yoshiyuki Nishinaka, Kouichi Kishida, and Yunwen Ye. Evolution patterns of open-source software systems and communities. In *Proceedings of the international workshop on Principles of software evolution IWPSE 02*, volume 2002 of *IWPSE '02*, pages 76–85. ACM Press, 2002.

[25] Kevin Nam, Mark S Ackerman, and Lada Adamic. Questions in, knowledge in?: a study of naver's question answering community. *CHI*, pages 779–788, 2009.

[26] Aditya Pal, Shuo Chang, and Joseph A Konstan. Evolution of Experts in Question Answering Communities. In *6th AAAI International Conference on Weblogs and Social Media*, ICWSM '12, pages 274–281, Dublin, Ireland, 2012. The AAAI Press.

[27] Aditya Pal and Scott Counts. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 45–54, New York, NY, USA, 2011. ACM.

[28] Aditya Pal, Rosta Farzan, Joseph A Konstan, and Robert E Kraut. Early Detection of Potential Experts in Question Answering Communities. *Human-Computer Interaction*, pages 231–242, 2011.

[29] Katherine Panciera, Aaron Halfaker, and Loren Terveen. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work*, GROUP '09, pages 51–60, New York, NY, USA, 2009. ACM.

[30] Katherine Panciera, Reid Priedhorsky, Thomas Erickson, and Loren Terveen. Lurking? cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1917–1926, New York, NY, USA, 2010. ACM.

[31] Fatemeh Riahi, Zainab Zolaktaf, Mahdi Shafiei, and Evangelos Milios. Finding expert users in community question answering. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 791–798, New York, NY, USA, 2012. ACM.

[32] Eduarda Mendes Rodrigues, Natasa Milic-Frayling, and Blaz Fortuna. Social Tagging Behaviour in Community-Driven Question Answering. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '08, pages 112–119, Washington, DC, USA, 2008. IEEE Computer Society.

[33] Yla R Tausczik and James W Pennebaker. Participation in an online mathematics community: differentiating motivations to add. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 207–216, New York, NY, USA, 2012. ACM.

[34] Tammara Combs Turner, Marc A Smith, Danyel Fisher, and Howard T Welser. Picturing Usenet: Mapping Computer-Mediated Collective Action. *Journal of Computer-Mediated Communication*, 10(4):0, 2005.

[35] J H Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[36] Howard T Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding social roles in Wikipedia. In *Methods*, volume 48 of *iConference '11*, pages 122–129. ACM, ACM, 2011.

[37] Howard T Welser, Eric Gleave, Danyel Fisher, and Marc Smith. Visualizing the Signatures of Social Roles in Online Discussion Groups. *Journal of Social Structure*, 8(2):1–32, 2007.

[38] Steve Whittaker, Loren Terveen, Will Hill, and Lynn Cherny. The dynamics of mass interaction. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, CSCW '98, pages 257–264, New York, NY, USA, 1998. ACM.

# Apêndice A

# Centros não normalizados da análise de perfis de longo prazo

Tabela A.1: Centros não normalizados dos grupos formados pelos perfis de longo prazo identificados na análise dos 36 sites.

| Perfis | Resp. | Perg. | Comen. | Tempo Atuação (# dias) | UMResp. | UMPerg. | UMComen. |
|---|---|---|---|---|---|---|---|
| Passageiro | 0,64 | 1,09 | 1,56 | 2,45 | -0,01 | 1,02 | 0,10 |
| Imperito em respostas | 1,65 | 0,65 | 1,21 | 2,75 | -0,86 | 0,39 | 0,07 |
| Ocasional | 6,96 | 4,67 | 16,54 | 12,70 | -0,03 | 9,16 | 0,36 |
| Expert em respostas | 1,86 | 1,07 | 3,02 | 3,87 | 1,14 | 1,05 | 0,16 |
| Expert em perguntas | 6,38 | 1,91 | 13,15 | 9,50 | 0,07 | 76,18 | 0,93 |
| Expert em comentarios | 1,70 | 0,41 | 2,19 | 3,56 | -0,02 | 1,55 | 8,92 |
| Ativista em Q-A | 64,98 | 30,31 | 161,08 | 88,84 | -0,04 | 4,89 | 0,53 |
| Ativista respondedor | 346,08 | 26,90 | 806,02 | 280,89 | 0,21 | 8,09 | 0,90 |
| Hiperativista | 1130,66 | 40,55 | 2813,60 | 558,59 | 0,28 | 11,41 | 0,98 |

# Apêndice B

# Composição e produção dos perfis nos 36 sites

| Site | Passageiro | Imperito em respostas | Ocasional | Expert em respostas | Expert em comentários | Ativista em Q-A | Ativista resp. + Hiperativista |
|---|---|---|---|---|---|---|---|
| Mathematics | 20,98 | -19,37 | -9,36 | -19,28 | -3,19 | 3,1 | 0,35 |
| Meta | 11,34 | -5,35 | -12,44 | -10,49 | 1,03 | 1,02 | 0,84 |
| Server Fault | 7,91 | -5,02 | -16,29 | 6,51 | -4,7 | 0,34 | -2,75 |
| Stats | 6,66 | -9,3 | 2,58 | -7,88 | -1,76 | 1,37 | 1,32 |
| Ask Ubuntu | 6,42 | -10,98 | 3,34 | -2,55 | -0,23 | -4,21 | -2,15 |
| TeX | 6,11 | -8,57 | 5,66 | -10,17 | -2,26 | -0,61 | 0,91 |
| Wordpress | 2,91 | -0,96 | -5,59 | -2,78 | 7,48 | 1,17 | 0,12 |
| DBA | 1,75 | -5,27 | 3,72 | -2,34 | 3,3 | 0,29 | 1,72 |
| Security | 1,18 | 2,17 | -4,35 | -1,08 | 0,38 | -2,1 | 0,57 |
| Electronics | 0,5 | -3,39 | 6,31 | -3,89 | 0,17 | -0,29 | 1,9 |
| GIS | -1,37 | -3,7 | 9,26 | -3,25 | 1,97 | 3,62 | -0,03 |
| Physics | -1,41 | -0,34 | 6,35 | -4,51 | -0,69 | 3,65 | 0,12 |
| Skeptics | -1,61 | -1,8 | 6,55 | -2,08 | 3,34 | 1,52 | 1,18 |
| WebMasters | -1,96 | -4,04 | 7,81 | 3,88 | 1,6 | -2,07 | -1,96 |

Figura B.1: (Parte-I) Resíduos obtidos no teste Chi-quadrado de Pearson para verificar a dependência das composições. Células coloridas representam resultados com uma diferença significativa ($p < ,005$), sendo a cor esverdeada uma diferença positiva e a avermelhada negativa.

| Site | Passageiro | Imperito em respostas | Ocasional | Expert em respostas | Expert em comentários | Ativista em Q-A | Ativista resp. + Hiperativista |
|---|---|---|---|---|---|---|---|
| DIY | -2,22 | -2,46 | 10,9 | -2,88 | 1,05 | -0,72 | 1,28 |
| CS Theory | -2,35 | -0,68 | 7,51 | -2,38 | 1,78 | 1,08 | 1,55 |
| Super User | -2,45 | 11,26 | -17,57 | 14,34 | -7 | -3,55 | -3,83 |
| SharePoint | -2,48 | -3,66 | 8,99 | -0,82 | 5,04 | 4,04 | 0,64 |
| GameDev | -2,58 | 2,15 | 2,83 | -0,13 | -3,56 | 2,84 | 2,15 |
| Drupal | -3,91 | -2,17 | 7,38 | 1,42 | 9,07 | 5,14 | 0,3 |
| Unix | -4,33 | 3,37 | 3,38 | 3,51 | 3,12 | -2,78 | 0,05 |
| Programmers | -4,38 | 10,09 | -9,64 | 5,61 | -1,02 | 4,7 | 3,08 |
| UX | -5,25 | 6,61 | 3,13 | 0,13 | 1,07 | 0,06 | 1,59 |
| WebApps | -5,54 | -0,21 | 8,71 | 7,1 | -0,54 | -1,67 | -0,32 |
| Android | -5,66 | -1,71 | 12,93 | 3,21 | 5,28 | -1,8 | 0,02 |
| Photo | -5,74 | 4,57 | 8,19 | -2,82 | 6,03 | 1,21 | 0,27 |
| English | -6,11 | 9,12 | 7,15 | -5,98 | 3,3 | -2,76 | 2,54 |
| Apple | -6,9 | 4,23 | 8,17 | 3,41 | 0,34 | -0,07 | 0,02 |
| Cooking | -8,99 | 8,02 | 10,96 | -1,07 | 3,1 | -0,34 | 1,6 |
| Sci-fi | -10,38 | 9,56 | 10,37 | 0,83 | 6,56 | -1,73 | 1,95 |
| Gaming | -16,21 | 18,11 | 12,24 | 3,77 | -3,69 | -0,02 | 1,22 |

Figura B.2: (Parte-II) Continuação da Figura B.1.

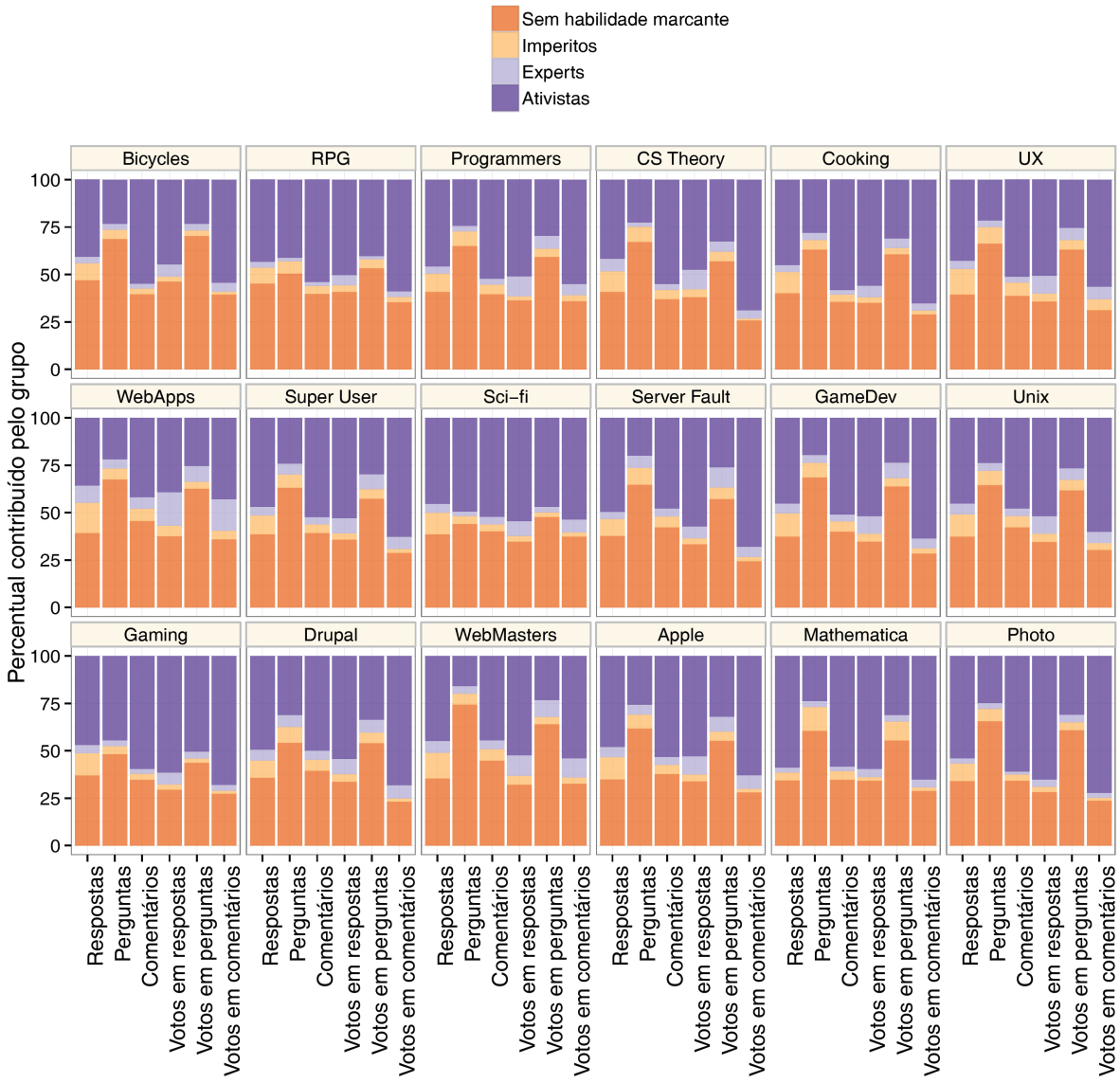


Figura B.3: (Parte-I) Produção dos grupos nos 36 sites. Os resultados estão ordenados pelo percentual de respostas de contribuidores *sem habilidade marcante*, grupo que se refere a usuários de perfil passageiro e ocasional.

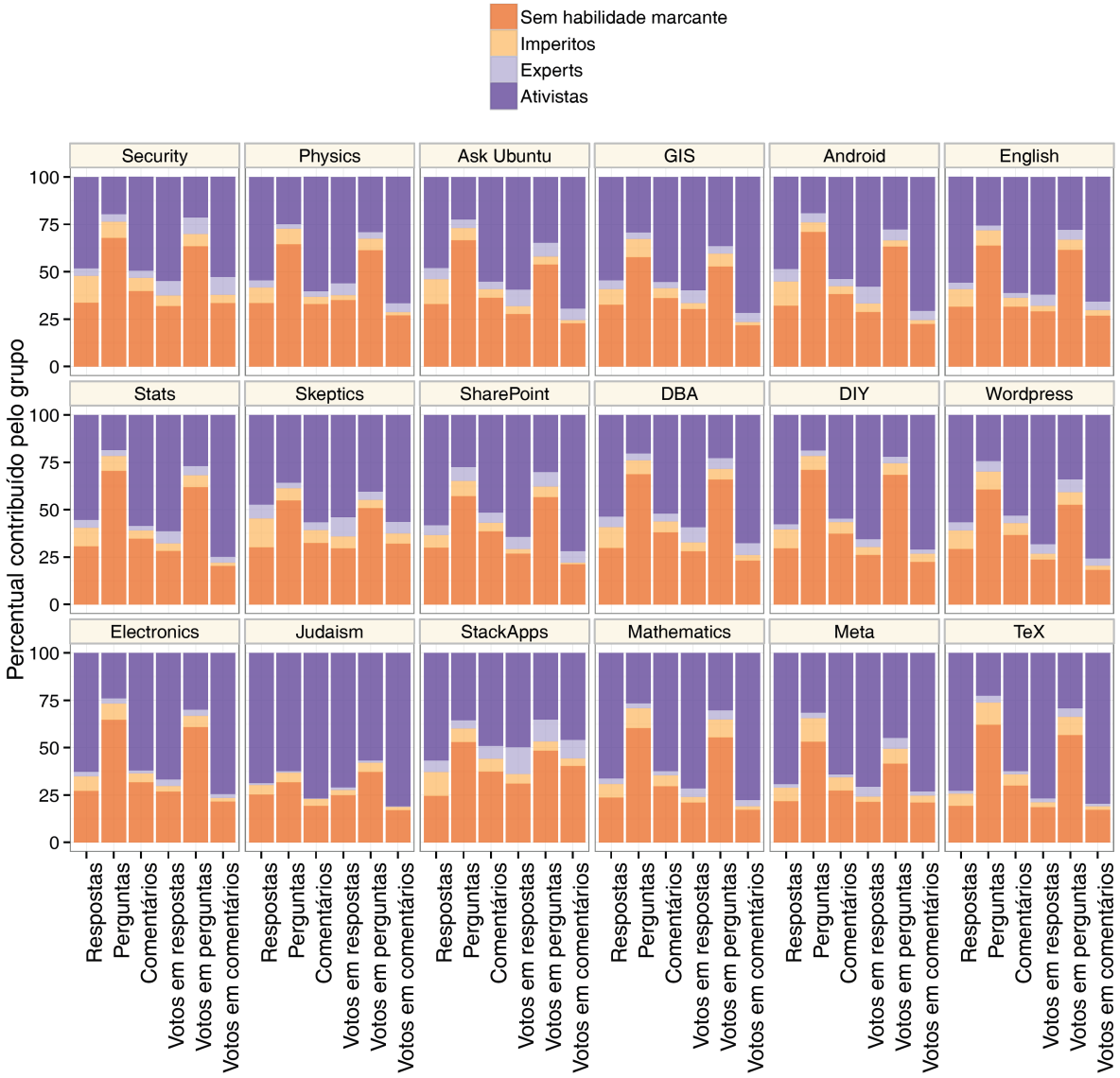


Figura B.4: (Parte-II) Continuação da Figura B.3.

# Apêndice C

# Resultados do agrupamento da análise de perfis de curto prazo

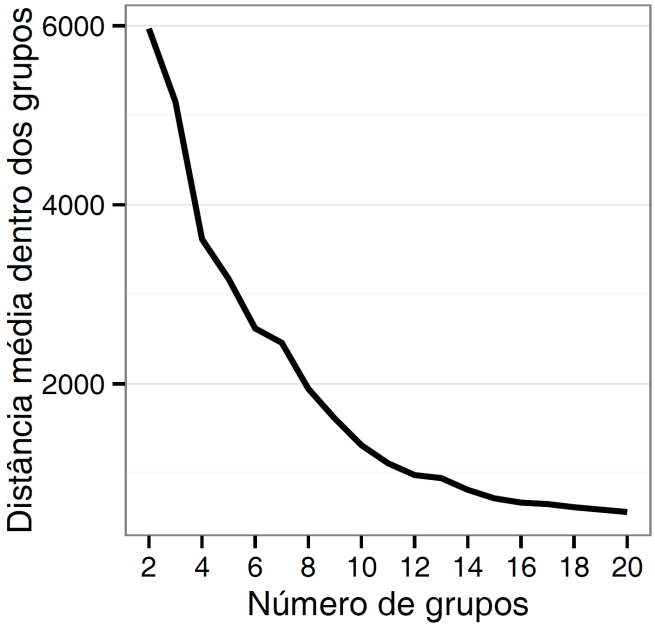


Figura C.1: Análise da heterogeneidade das soluções do agrupamento hierárquico no Super User.

Tabela C.1: Centros normalizados dos grupos identificados no agrupamento hierárquico no Super User. A primeira parte da tabela lista os centros da solução com 9 grupos, e a segunda lista os novos centros que surgem nas soluções de 10 e 11 grupos.

| Centros | Resp. | Perg. | Comen. | Tempo Atuação | UMResp. | UMPerg. | UMComen. |
|---|---|---|---|---|---|---|---|
| Centro 1 | -0.19 | 0.17 | -0.13 | -0.23 | 0.05 | 0.00 | -0.19 |
| Centro 2 | 0.92 | -0.05 | 0.67 | 1.81 | -0.01 | -0.11 | 0.27 |
| Centro 3 | -0.07 | -0.35 | -0.17 | -0.27 | -1.37 | -0.33 | -0.22 |
| **Centro 4*** | **-0.15** | **-0.52** | **-0.17** | **-0.36** | **0.07** | **-0.43** | **-0.23** |
| Centro 5 | -0.15 | 0.08 | -0.07 | -0.16 | -0.07 | 2.30 | -0.13 |
| **Centro 6**** | **-0.09** | **-0.40** | **-0.08** | **-0.09** | **-0.10** | **-0.31** | **2.41** |
| Centro 7 | -0.07 | -0.19 | -0.12 | -0.15 | 2.06 | -0.20 | -0.09 |
| Centro 8 | 0.02 | 3.86 | 0.57 | 1.44 | -0.17 | 0.33 | -0.06 |
| Centro 9 | 6.77 | 0.92 | 6.22 | 6.61 | 0.23 | 0.36 | 0.42 |
| **Novo Centro - 10*** | **-0.20** | **0.02** | **-0.16** | **-0.32** | **0.03** | **-0.22** | **-0.23** |
| **Novo Centro - 11**** | **-0.11** | **-0.41** | **-0.07** | **-0.11** | **-0.23** | **-0.36** | **1.22** |

Tabela C.2: Centros não normalizados dos grupos identificados no Super User.

| Perfis | Resp. | Perg. | Comen. | Tempo Atuação (# dias) | UMResp. | UMPerg. | UMComen. |
|---|---|---|---|---|---|---|---|
| Passageiro | 0,79 | 0,65 | 1,50 | 1,92 | 0,01 | 0,37 | 0,10 |
| Imperito em respostas | 1,84 | 0,26 | 1,27 | 2,14 | -0,84 | 0,23 | 0,07 |
| Ocasional | 1,00 | 2,10 | 3,76 | 3,55 | -0,01 | 4,01 | 0,10 |
| Expert em respostas | 1,88 | 0,51 | 2,10 | 2,73 | 1,15 | 0,53 | 0,13 |
| Expert em perguntas | 2,56 | 1,37 | 6,55 | 4,06 | 0,09 | 31,32 | 0,41 |
| Expert em comentarios | 1,11 | 0,27 | 2,14 | 2,53 | 0,03 | 0,71 | 6,71 |
| Ativista em Q-A | 4,51 | 12,06 | 21,41 | 13,85 | -0,04 | 2,35 | 0,11 |
| Ativista respondedor | 38,38 | 1,61 | 54,36 | 27,83 | 0,10 | 1,59 | 0,57 |
| Hiperativista | 191,39 | 3,44 | 322,53 | 52,77 | 0,18 | 3,27 | 0,51 |

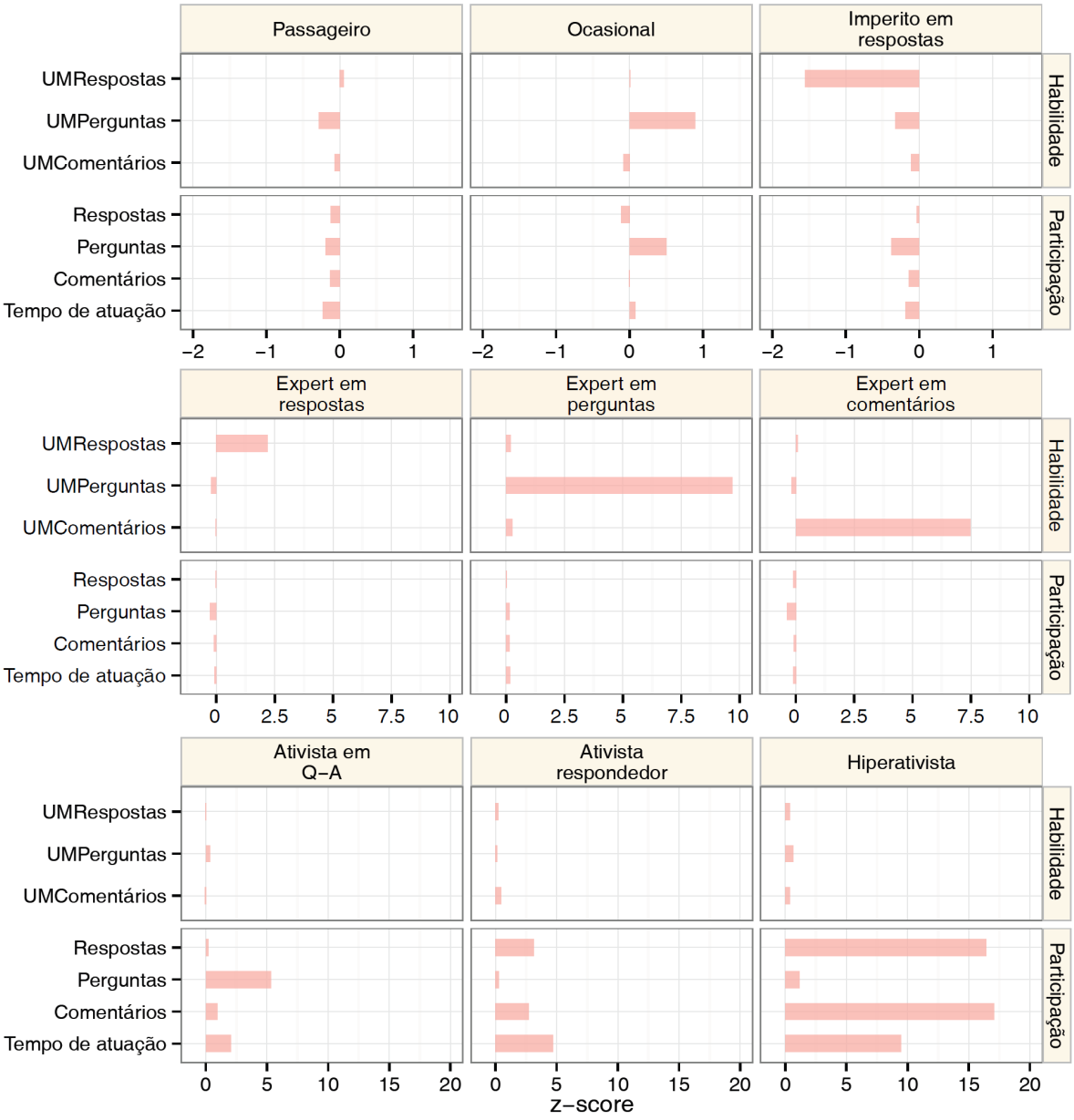


Figura C.2: Centros dos nove grupos identificados na análise de agrupamento de 15 janelas de 2 meses do Super User. Note que o eixo horizontal é o z-score da métrica e as escalas nos três grupos de gráficos variam.

# Apêndice D

# Tutorial para a reprodução do experimento

Neste apêndice descrevemos um passo-a-passo de como executar todas as etapas do nosso experimento de análise de perfis. O tutorial apresentado a seguir detalha as ferramentas utilizadas, a localização dos dados, códigos e *scripts*, e informações de como executá-los.

## D.1 Ferramental

Lista das ferramentas necessárias para a reprodução do experimento:

- Ruby 1.9.3-p327 (>=) com as bibliotecas: libxml e mysql;
- MySQL Server 5.5 (>=);
- Java 7 (>=);
- R 2.15.2 (>=) com as bibliotecas: cluster, ggplot2, reshape, traminer, gmodels e mass.

## D.2 Configuração do ambiente

Etapas necessárias para a configuração do ambiente:

1. Baixar o projeto no repositório SVN[1] - *https://svn.lsd.ufcg.edu.br/repos/sc/StackExchange/trunk*;

[1] Checkout como usuário anonymous

2. Baixar a base dos dados lançada no mês de agosto de 2012 - *http://www.clearbits.net/torrents/2076-aug-2012*. Esses dados estão no formato XML e contêm a atividade dos sites desde sua criação até o dia 31 de julho de 2012;

3. Criar um banco no MySQL para cada site e carregar seu esquema;

    - O SQL do esquema se encontra na pasta *scripts/mysql* do projeto.

4. Executar o script Ruby que carrega os dados em XML para o banco MySQL.

    - O script *load* se encontra na pasta *scripts/mysql* do projeto.

## D.3 Execução do experimento

Informações sobre os códigos e scripts que devem ser executados para processar os dados dos sites de Stack Exchange.

### D.3.1 Extração das métricas de participação e habilidade

1. Executar o código do extrator das métricas (Java) - *StackExchange-Extractor*.

    - Para análise de perfis nos 36 sites - *ExtractorStatic.java*;
    - Para análise da dinâmica dos perfis - *ExtractorDynamic.java*.

### D.3.2 Análise de agrupamento

1. Executar o script R que realiza o agrupamento dos dados;

    - Para análise de perfis nos 36 sites - scripts/allsites/*agrupamento.R*;
    - Para análise da dinâmica dos perfis - scripts/n_janelas/*agrupamento.R*.

2. Executar o script R que produz o gráfico dos centros dos grupos.

    - Para análise de perfis nos 36 sites - scripts/allsites/grafico/*centro_dos_grupos.R* e scripts/allsites/grafico/*centros_scatterplot.R*;
    - Para análise da dinâmica dos perfis - scripts/n_janelas/grafico/*centro_dos_grupos.R*.

### D.3.3 Análise de composição e produção dos perfis nos 36 sites

1. Executar o script R que processa os resultados da análise de agrupamento, e gera os gráficos de composição e produção dos perfis.

    - Caminho: scripts/allsites/*composicao_producao.R*.

### D.3.4 Análise de dinâmica dos perfis no Super User

1. Executar o script R que processa os resultados da análise de agrupamento, e produz os dados de entrada dos gráficos da análise de dinâmica;

    - Caminho: scripts/n_janelas/*gera_entrada_dinamica.R*.

2. Executar o script R que gera os gráficos da análise de dinâmica.

    - Caminho: scripts/n_janelas/*graficos_dinamica.R*.