Como Pesquisadores Usam o Dataset GHTorrent?

Hudson Borges, Jailton Coelho, Paulo Carvalho
Mariane Fernandes, Marco Tulio Valente

[1] ASERG Group – Departamento de Ciência da Computação (DCC)
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – Minas Gerais – Brasil.

{hsborges,jailtoncoelho,paulocarvalho,marianef,mtov}@dcc.ufmg.br

Abstract. *GitHub's public data has been extensively used as data source in recent studies. However, little is known about how researchers actually use this data. In this paper, we present an exploratory study on how researchers use this data. For this purpose, we analyze 114 scientific articles that use GHTorrent, the main GitHub dataset. We show that repositories and users data are very often exploited in these studies. We also observed that the amount of data used varies considerably and only a small fraction of the data available is used.*

Resumo. *Dados públicos do GitHub tem sido exaustivamente utilizados como fonte de pesquisa em estudos recentes. Contudo, pouco se sabe como pesquisadores efetivamente usam esses dados. Este trabalho apresenta um estudo exploratório que investiga como pesquisadores usam tais dados. Para isso, foram analisados 114 artigos científicos que usam o principal dataset de dados do GitHub, o GHTorrent. Neste estudo, mostra-se que dados de repositórios e usuários são muito frequentemente explorados. Observou-se também que a quantidade de dados utilizados varia consideravelmente e somente uma pequena fração dos dados disponíveis no dataset é utilizada.*

1. Introdução

Com aproximadamente 22 milhões de usuários e 60 milhões de repositórios, GitHub é a plataforma de hospedagem de código, controle de versão e colaboração mais popular atualmente.[1] Nos últimos anos, GitHub tem sido amplamente utilizado por pesquisadores como fonte de pesquisa, principalmente, por oferecer acesso a dados públicos por meio de sua Web API (*Application Programming Interface*). Contudo, a coleta dos dados por meio dessa API não é trivial e está sujeita a diversas restrições. Por exemplo, usuários possuem um limite máximo de requisições que podem ser realizados por hora. Além disso, para completa obtenção de dados, muitas vezes é necessário o acesso a diferentes pontos da API [Gousios and Spinellis 2012].

Com intuito de faciliar o acesso a esses dados por desenvolvedores e pesquisadores, alguns *datasets* foram propostos, como GHTorrent [Gousios and Spinellis 2012, Gousios 2013], GitHub Archive [Grigorik 2012] e BOA [Dyer et al. 2013]. Estudos recentes mostram que, depois da própria API do GitHub, o GHTorrent é a fonte de dados mais utilizada em pesquisas na área de Engenharia de Software [Cosentino et al. 2016].

[1] `https://github.com/search`, verificado em 04/07/2017.

Gousios e Spinellis (2012) definem o GHTorrent como um serviço de coleta de dados públicos do GitHub que tem por objetivo facilitar o acesso aos dados oferecidos pela API oficial por desenvolvedores e pesquisadores. Em linhas gerais, esse serviço monitora os eventos lançados pela API oficial, coleta informações adicionais necessárias e os disponibiliza diretamente (*i.e.*, no mesmo formato recebido pela API) ou de forma estruturada (*i.e.*, em um banco de dados relacional).

Embora dados públicos do GitHub sejam exaustivamente explorados por pesquisadores, pouco se sabe sobre quais informações e a quantidade de dados comumente utilizada em suas pesquisas. Assim, neste artigo apresenta-se um estudo exploratório que investiga como pesquisadores efetivamente usam o *dataset* mais popular de dados do GitHub, o GHTorrent. Especificamente, são propostas duas questões de pesquisa centrais:

QP#1. Quais são os usos mais comuns do *dataset* GHTorrent?

QP#2. Qual a quantidade de dados utilizados por estudos baseados no GHTorrent?

Para responder a essas questões, foram analisados 254 artigos científicos que citam os dois principais trabalhos dos autores do GHTorrent. Assim, as principais contribuições deste artigo incluem a identificação de quais dados oferecidos pelo GHTorrent são mais utilizados, assim como a quantidade de dados que são utilizados por pesquisadores.

O restante deste artigo está organizado conforme descrito a seguir. Na Seção 2 é apresentada a metodologia de pesquisa adotada e na Seção 3 os resultados obtidos. Na Seção 4 os riscos à validade são apresentados e na Seção 5 são discutidos trabalhos relacionados. Por fim, na Seção 6 são apresentadas as conclusões e trabalhos futuros.

2. Metodologia do Estudo

Como discutido na seção anterior, os objetivos deste trabalho são: (i) identificar quais são os usos mais comuns dos dados oferecidos pelo GHTorrent, e (ii) avaliar o quanto dos dados disponíveis no *dataset* é efetivamente utilizado por pesquisadores. Na Seção 2.1 é detalhado como foram coletados os trabalhos analisados neste estudo e na Seção 2.2 são apresentadas decisões e detalhes do processo de análise dos mesmos.

2.1. Coleta de Artigos

Para identificar como pesquisadores usam os dados do GHTorrent, foram analisados os artigos que citaram os trabalhos de Gousios e Spinellis (2012) e Gousios (2013). Estes dois trabalhos foram selecionados como fontes de pesquisa por dois motivos: (a) o primeiro trabalho introduziu o conceito, a arquitetura e a primeira implementação do *dataset*; e (b) o segundo artigo foi selecionado pois, no *website* do GHTorrent, os autores pedem explicitamente para citá-lo caso o *dataset* seja utilizado.[2] A identificação dos artigos que citam estes dois trabalhos foi feita por meio da ferramenta Google Scholar[3] e somente trabalhos escritos em português e inglês foram considerados neste estudo. Além disso, também foram excluídos desse estudo artigos nos quais os textos não estavam disponíveis ou eram rascunhos que foram publicados posteriormente. Na Tabela 1 são apresentados detalhes da coleta dos trabalhos que citaram o GHTorrent. Embora o primeiro trabalho, *i.e.* Gousios e Spinellis (2012), tenha introduzido o *dataset*, o segundo trabalho, *i.e.*

[2] `http://ghtorrent.org/`

[3] `https://scholar.google.com.br/`

Gousios (2013), concentra o maior número de citações, isto é, 174 de 254 trabalhos analisados. Além disso, 25 artigos fazem referência a ambos trabalhos e 32 foram excluídos pelos motivos descritos anteriormente. Portanto, neste estudo foi analisado um conjunto de 197 artigos científicos que incluem uma citação ao GHTorrent.

Tabela 1. Sumários dos trabalhos analisados

Fonte	Citações	Interseção	Removidos	Analisados
Gousios e Spinellis (2012)	80	25	15	65
Gousios (2013)	174		17	132
Total	254	25	32	197

2.2. Análise dos Artigos

Para a caracterização do uso do *dataset*, foram considerados os tipos de dados oferecidos pelo GHTorrent. Especificamente, foram consideradas as entidades descritas no modelo relacional do *dataset*. Neste modelo, as entidades representam informações de elementos do GitHub (*e.g.*, Repositório, Usuário e *Issue*) e são consistentes com os diferentes pontos de acesso da API oficial. Assim, para cada artigo é possível que mais de uma entidade seja utilizada. Por exemplo, o trecho abaixo foi extraído do trabalho de Kalliamvakou et al. (2014) e reporta um estudo sobre repositórios e usuários baseado em dados obtidos do GHTorrent.

"*We document the results of an empirical study aimed at understanding the characteristics of the repositories and users in GitHub ... and we provide evidence of these perils based on quantitative analysis of the GHTorrent dataset*". (Kalliamvakou et al., 2014).

Contudo, nem todos os trabalhos efetivamente usam o GHTorrent como fonte de dados em suas pesquisas. Durante a análise dos artigos, foram identificados trabalhos que citam os trabalhos originais do GHTorrent como trabalhos relacionados ou fonte alternativa de dados, por exemplo. Na Tabela 2 são apresentados detalhes do número de artigos que usam o GHTorrent apenas como referência bibliográfica e os que usam como fonte de dados. Observa-se que a maioria dos artigos, 114 de 197, usam o GHTorrent como fonte de dados. Observa-se também que existe um comportamento inverso entre os dois trabalhos considerados, sendo o primeiro mais utilizado como referência bibliográfica e o segundo como fonte efetiva de dados para estudos e experimentos.

Tabela 2. Sumário do tipo de uso do dataset GHTorrent

Trabalho	Uso	
	Referência	Fonte de dados
Gousios e Spinellis (2012)	36	29
Gousios (2013)	47	85
Total	83	114

Por fim, para responder a QP #2, foram considerados somente os números absolutos de dados utilizados e reportados pelos autores. Além disso, em alguns casos, o total

de dados utilizado pelos pesquisadores não foi devidamente reportado. Por exemplo, no trabalho de Robinson e Deng (2015) é descrito que são utilizados dados sobre projetos, *issues* e *pull requests*, contudo os autores informam, sem mais detalhes, que tais dados foram obtidos de 103 repositórios. Portanto, em casos como este, somente os valores reportados são utilizados para responder à questão de pesquisa.

"*We analyzed 103 projects from GitHub, the most popular open-source code repository site ... Our data was derived mainly from these collections: issues, issue events, issues comments, pull requests, and pull request comments.*" (Robinson e Deng, 2015).

3. Resultados

QP #1. Quais são os usos mais comuns do dataset GHTorrent?

Nessa primeira questão de pesquisa estuda-se os usos mais comuns dos dados disponibilizados pelo *dataset* GHTorrent. Na Figura 1 são apresentadas as entidades que foram mais utilizadas nos trabalhos analisados.

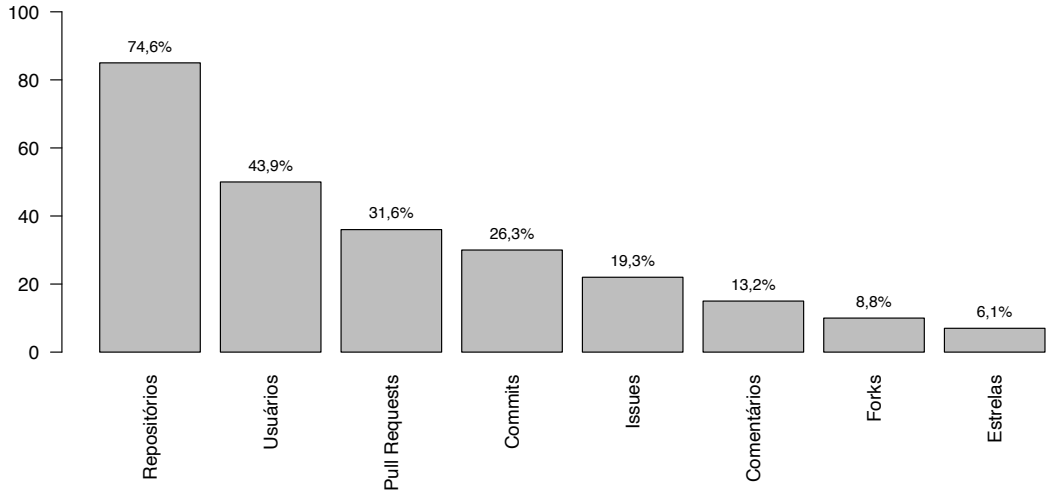


Figura 1. Dados mais utilizados do dataset GHTorrent

A seguir, cada uma das entidades são apresentadas em detalhes.

Repositórios: Dos 114 trabalhos analisados, 85 (74,6%) reportaram o uso de dados relacionados a repositórios. A relevância dessa entidade para pesquisadores está diretamente relacionado à centralidade que repositórios possuem no modelo de dados do GHTorrent. Em geral, repositórios são utilizados como ponto de partida para seleção de candidatos a inclusão em estudos. Por exemplo, é comum que selecionem somente repositórios com um número mínimo ou maximo de estrelas, *commits*, *issues*, etc.

Usuários: De forma similar a repositórios, usuários possuem um papel central no modelo de dados. Especificamente, usuários são responsáveis por ações e eventos no GitHub (*e.g.*, criação de *commits*, *issues* e *pull requests*). Neste estudo, identificou-se que 49 dos 114 trabalhos (43,9%) reportaram o uso de dados de usuários.

Pull Requests, Commits e Issues: *Pull requests*, *commits* e *issues* são registros de atividades relacionadas ao desenvolvimento de software e foco de diversos estudos recentes. Neste estudo, foram identificados 36 trabalhos (31,6%) que mencionam o uso de dados relacionados a *pull requests*. 30 trabalhos (26,3%) reportaram o uso de dados relacionados a *commits*. Além disso, 22 (19,3%) relataram o uso de dados sobre *issues*.

Comentários: Como uma plataforma social de desenvolvimento, GitHub permite discussões entre desenvolvedores por meio de comentários em *issues*, *pull requests*, *commits*, etc. Foram identificados 15 trabalhos (13,2%) que usam tais tipos de comentários em suas investigações.

Forks: Embora forks sejam um subconjunto de repositórios, esses foram explicitamente mencionados como fonte de pesquisa por 10 trabalhos (8,8%).

Estrelas: No GitHub, estrelas é uma funcionalidade semelhante ao "*like*" em outras plataformas sociais (*e.g.*, YouTube e Facebook) e frequentemente usada como medida de popularidade. Neste estudo, foram identificados 7 trabalhos (6,1%) que usam tais dados.

Na Tabela 3 são apresentadas as top-10 combinações de entidades do GHTorrent que foram frequentemente utilizadas nos artigos analisados. Observa-se que a maioria dos trabalhos incluem dados de repositórios e usuários. De fato, dados relacionados a repositórios ou a usuários estão presentes em 9 das 10 combinações mais frequentes, evidenciando a importância destes dados para pesquisadores. Observa-se também que o uso de dados de repositórios com *pull requests* (29 vezes), *commits* (23 vezes) e *issues* (16 vezes) é bastante frequente. Nestes casos, o uso de repositórios é feito como forma de seleção de candidatos e correlação com perfis de projetos. Outras combinações comumente utilizadas envolvem o uso de dados de usuários com *commits* (16 vezes), *pull requests* (15 vezes) e *issues* (13 vezes). Constatou-se que trabalhos que usam esses dados buscam relacionar as atividades com os perfis dos usuários que as realizaram. Por fim, em 11 trabalhos, observou-se o uso de dados de *commits* e *pull requests*, indicando o interesse nas modificações que são feitas antes do envio para o projeto de origem.

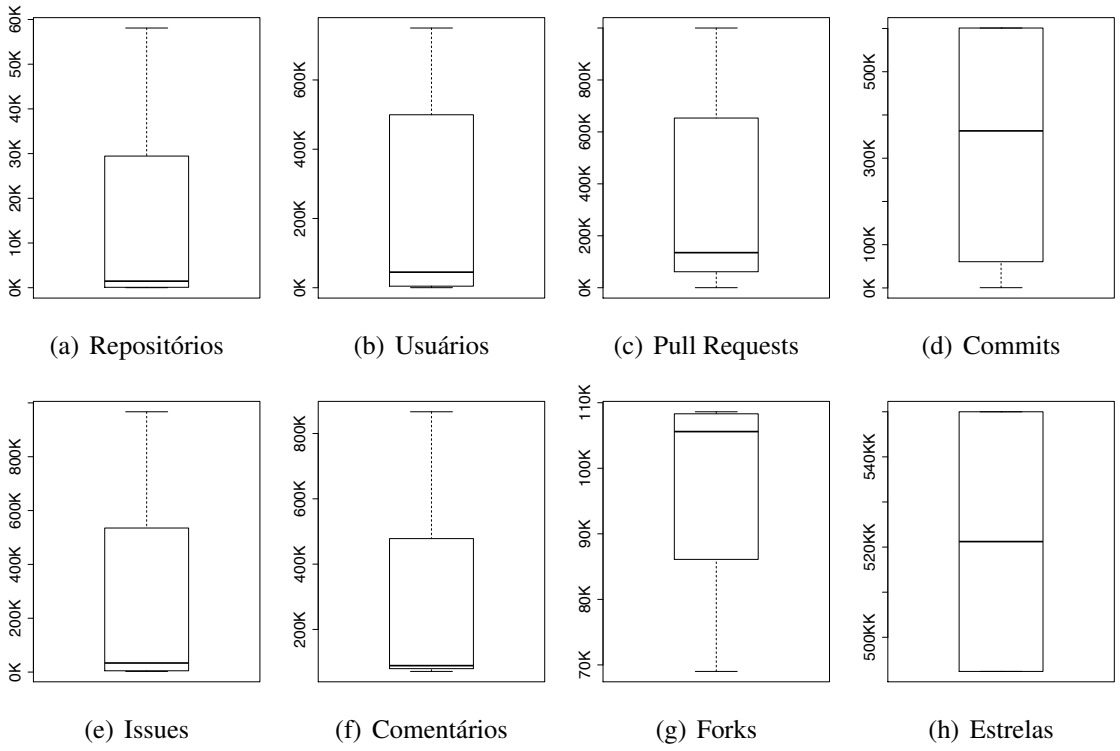
Tabela 3. Top-10 combinações de entidades do GHTorrent mais frequentes

Entidades	Freq.	Entidades	Freq.
Repositórios + Usuários	33	Usuários + Pull Requests	15
Repositórios + Pull Requests	29	Usuários + Issues	13
Repositórios + Commits	23	Repositórios + Usuários + Pull Requests	13
Repositórios + Issues	16	Repositórios + Usuários + Commits	13
Usuários + Commits	16	Commits + Pull Requests	11

QP #2. Qual a quantidade de dados utilizados por estudos baseados no GHTorrent?

Nessa segunda questão de pesquisa analisa-se a quantidade de dados utilizados pelos pesquisadores. Na Figura 2 são apresentadas as distribuições das quantidades de dados utilizada por entidade. No total, 74 trabalhos reportaram o uso de 7 até 34,6 milhões de repositórios. Os valores para o primeiro, segundo e terceiro quartis são 90, 1.469 e 29.447 repositórios, respectivamente. Para usuários, foi identificada uma variação entre 150 e 13,2 milhões de dados utilizados em 45 trabalhos. Os valores para o primeiro,

segundo e terceiro quartis são 4.500, 45.003 e 499.485 usuários, respectivamente. O total de *pull requests* utilizados foi reportado por 18 trabalhos. Esse número varia entre 218 e 3,2 milhões e os valores para o primeiro, segundo e terceiro quartis são 61.592, 135.084 e 1 milhão, respectivamente. Considerando o total de *commits*, foi identificada uma variação entre 600 e 1,5 bilhão. Neste caso, os valores para o primeiro, segundo e terceiro quartis são 60.658, 363.162 e 601.080 *commits*, respectivamente. Para *issues*, 11 trabalhos indicaram um total que varia entre 2.401 e 24,1 milhões. Os valores para o primeiro, segundo e terceiro quartis são 4.790, 33.673 e 535.160 *issues*, respectivamente. Por fim, quatro trabalhos reportaram o número de comentários e forks utilizados, sendo a mediana de 59.132 e 105.596, respectivamente. Além disso, dois trabalhos utilizaram um total de 55 milhões e 49,2 milhões de estrelas disponíveis.



(a) Repositórios (b) Usuários (c) Pull Requests (d) Commits

(e) Issues (f) Comentários (g) Forks (h) Estrelas

Figura 2. Número total de dados utilizados por entidade (*outliers foram omitidos*)

4. Ameaças à Validade

Validade Externa: Este estudo limitou-se à analise de 114 trabalhos que efetivamente usaram o GHTorrent como fonte de dados. Portanto, os resultados reportados não podem ser generalizados para todas as fontes de dados do GitHub, como a API oficial ou o *dataset* GitHub Archive. Entretanto, estudos recentes mostram que, depois da API oficial, o GHTorrent é o *dataset* mais utilizada por pesquisadores.

Validade Interna: Dentre os aspectos que podem afetar os resultados apresentados, destaca-se o método para identificação e coleta dos trabalhos analisados. A fim de minimizar essa ameaça, utilizou-se a função "citado por" da ferramenta Google Scholar.

Validade de Construção: Para a análise do uso do GHTorrent, foram analisados 254 artigos que citam os dois principais trabalhos científicos dos autores do *dataset*. Como discutido, esses dois trabalhos foram selecionados como fonte de pesquisa devido sua relevância e número de citações obtidas desde suas respectivas publicações. Contudo, os autores possuem outros trabalhos que também envolvem o uso do GHTorrent e podem ter sido usado como referência por outros pesquisadores.

5. Trabalhos Relacionados

Devido a sua popularidade, GitHub tem sido utilizado em estudos de diversas áreas, como popularidade de projetos [Borges et al. 2016a, Borges et al. 2016b], ecossistemas de software [Matragkas et al. 2014, Blincoe et al. 2015] e comportamentos sociais no desenvolvimento de software [Yu et al. 2014]. Em particular, este estudo foi motivado pelo trabalho de Cosentino et al. (2016). Neste trabalho, os autores analisam 93 artigos científicos sobre três aspectos: (i) o método empírico adotado, (ii) o *dataset* utilizado e (iii) as limitações reportadas pelos autores em seus trabalhos. Os resultados deste trabalho mostram que: (i) a maioria dos trabalhos que usam informações do GitHub se baseiam na observação direta dos dados, (ii) GHTorrent é o *dataset* mais usado nos trabalhos analisados e (iii) somente uma pequena fração dos trabalhos reportam ameaças relacionadas ao *dataset* utilizado. Embora neste trabalho os autores tenham identificado as principais fontes de dados e métodos de análise dos dados, os mesmos não investigaram os usos que foram feitos dessas fontes. No estudo apresentado neste artigo, detalhamos como pesquisadores usam o principal *dataset* de dados do GitHub, o GHTorrent.

Outros trabalhos analisam características de repositorios *git* com objetivo de identificar as vantagens e os perigos de usar tais informações [Bird et al. 2009, Kalliamvakou et al. 2014, Kalliamvakou et al. 2016]. Nesses trabalhos, os autores fornecem recomendações e alertas a pesquisadores ao utilizar tais fontes de pesquisa em seus trabalhos. Por outro lado, no presente estudo, apresentamos uma investigação sobre como pesquisadores efetivamente usam o *dataset* mais popular de dados do GitHub.

6. Conclusão

Este trabalho apresentou um estudo sobre os usos mais comuns do *dataset* mais popular de dados do GitHub, o GHTorrent. Para isso, foram analisados 254 artigos que citaram os principais trabalhos científicos dos autores desse *dataset*. Na primeira questão de pesquisa, investigou-se quais são os dados mais utilizados por pesquisadores assim como os subconjuntos de dados mais utilizados. Além de centrais no modelo de dados, mostrou-se que repositórios e usuários são frequentemente utilizados, seja isoladamente, em conjunto, ou com outros dados. Outros dados frequentemente utilizados incluem *pull requests*, *commits*, *issues* e comentários. Na segunda questão de pesquisa, analisou-se a quantidade de dados utilizada pelos pesquisadores nos mesmos trabalhos. Verificou-se que a quantidade de dados pode variar consideravelmente, contudo pesquisadores usam somente uma pequena fração dos dados disponíveis. Por exemplo, metade dos trabalhos usam no máximo 1.469 repositórios. Como trabalhos futuros, pretende-se estender este estudo para outros *datasets* e verificar as semelhanças e diferenças no uso de ambos. Pretende-se também verificar se o uso dos dados segue as recomendações feitas por Bird et al. (2009) e Kalliamvakou et al. (2014) ao analisarem as possibilidades e perigos de se minerar repositórios *git*.

Agradecimentos: Esta pesquisa é financiada pela FAPEMIG, CAPES e pelo CNPq.

Referências

Bird, C., Rigby, P. C., Barr, E. T., Hamilton, D. J., German, D. M., and Devanbu, P. (2009). The promises and perils of mining git. In *6th Working Conference on Mining Software Repositories (MSR)*, pages 1–10.

Blincoe, K., Harrison, F., and Damian, D. (2015). Ecosystems in GitHub and a method for ecosystem identification using reference coupling. In *12th Working Conference on Mining Software Repositories (MSR)*, pages 202–207.

Borges, H., Hora, A., and Valente, M. T. (2016a). Predicting the popularity of GitHub repositories. In *12th International Conference on Predictive Models and Data Analytics in Software Engineering (PROMISE)*, page 9.

Borges, H., Hora, A., and Valente, M. T. (2016b). Understanding the factors that impact the popularity of GitHub repositories. In *32nd International Conference on Software Maintenance and Evolution (ICSME)*, pages 1–11.

Cosentino, V., Izquierdo, J. L. C., and Cabot, J. (2016). Findings from GitHub: methods, datasets and limitations. In *13th International Conference on Mining Software Repositories (MSR)*, pages 137–141.

Dyer, R., Nguyen, H. A., Rajan, H., and Nguyen, T. N. (2013). Boa: A language and infrastructure for analyzing ultra-large-scale software repositories. In *35th International Conference on Software Engineering (ICSE)*, pages 422–431.

Gousios, G. (2013). The GHTorrent dataset and tool suite. In *10th Working Conference on Mining Software Repositories (MSR)*, pages 233–236.

Gousios, G. and Spinellis, D. (2012). GHTorrent: GitHub's data from a firehose. In *9th Working Conference of Mining Software Repositories (MSR)*, pages 12–21.

Grigorik, I. (2012). The GitHub archive.

Kalliamvakou, E., Gousios, G., Blincoe, K., Singer, L., German, D. M., and Damian, D. (2014). The promises and perils of mining GitHub. In *11th Working Conference on Mining Software Repositories (MSR)*, pages 92–101.

Kalliamvakou, E., Gousios, G., Blincoe, K., Singer, L., German, D. M., and Damian, D. E. (2016). An in-depth study of the promises and perils of mining GitHub. *Empirical Software Engineering*, 21(5):2035–2071.

Matragkas, N., Williams, J. R., Kolovos, D. S., and Paige, R. F. (2014). Analysing the'biodiversity'of open source ecosystems: the GitHub case. In *11th Working Conference on Mining Software Repositories (MSR)*, pages 356–359.

Robinson, W. N. and Deng, T. (2015). Data mining behavioral transitions in open source repositories. In *48th Hawaii International Conference on System Sciences*, pages 5280–5289.

Yu, Y., Yin, G., Wang, H., and Wang, T. (2014). Exploring the patterns of social behavior in GitHub. In *1st Workshop on Crowd-based Software Development Methods and Technologies (CrowdSoft)*, pages 31–36.