Analyzing Government Use of GitHub for Collaboration: An Empirical Approach to Measuring Open Government and Open Collaboration

By

Jaydeep Mistry

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Environmental Studies
in
Geography

Waterloo, Ontario, Canada, 2020

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The way government organizations collaborate on developing computer software has significantly changed with the use of the Internet. GitHub, an online platform that hosts computer software and provides project management solutions, has been popular for hosting open source software projects. Although some government organizations have been adopting the use of GitHub for their own work, there is a lack of understand as to why they use it and how it can contribute to them becoming an open government. This research identifies motivations and challenges that they face in using the platform to become an open government, and how they are participating in open collaboration on the platform.

Governments are motivated to use GitHub because it allows them to break down silos of knowledge within government departments and share knowledge more freely. It comes with the challenges to train government workers to use version control systems such as Git, or to work within loose legal frameworks of what software is appropriate for governments to become an open government. As for the usage of the government accounts on the platform, almost 50% of government accounts on GitHub have actively used the platform since 2018. Although there are over 700 government organization accounts on GitHub, there is a lack of metadata or information available on their account as only 47% of them have provided a description about themselves, and only 36% have provided an email address to contact. Additionally, only 3% of all government accounts are verified accounts on GitHub.

There is a collaborative relationship between government accounts who use GitHub, however there is a long-tail distribution in the number of collaborations (node degree). Few government accounts such as @alphagov (United Kingdom), @18F (United States of America), or @govau (Australia) are the most frequent collaborators, and they are their respective country's chief open government organizations. Overall, this research demonstrates how to study the progression of open government and open collaboration using GitHub data, users, and organizations as a case study.

Acknowledgements

I thank my professor and advisor Dr. Peter Johnson for providing me with the opportunity to be in the Master of Environmental Studies program at the Department of Geography and Environmental Management, University of Waterloo. I also thank him, and all others at Faculty of Environment who have helped me get to where I am today.

I thank all the mentors that I have had the luck and privilege to work with throughout the years since starting my Master program. With your support and guidance, I have learned a great deal about work, friendship, and life.

I thank Philip Everson for running Velocity Start events. Your guidance was most inspiring to me to become more entrepreneurial. I give my condolences to your family for your unfortunate passing in 2019.

I also give my condolences to the family of Marzieh Foroutan for her unfortunate passing in 2020.

Dedication

I dedicate this thesis to my parents. With their support, guidance, and love I was able to accomplish all that I have done so far, and I look ahead for greater things to come.

Table of Contents

List of Figures

List of Tables

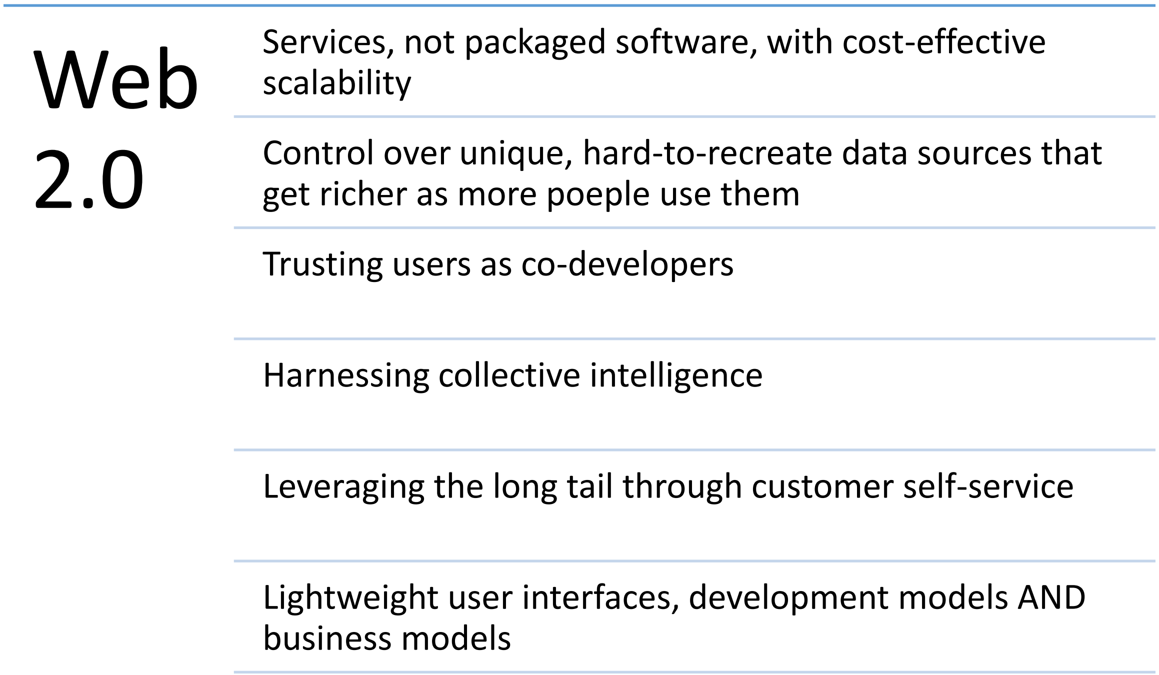
Chapter 1

1 Thesis Overview

1.1 Introduction

The way people collaborate with each other has changed significantly in the past few decades with the introduction of the Internet in the 1970s (Leiner, et al., 2009; Mowery & Simcoe, 2005). Before the dot-com bubble burst in 2001, users were mainly consumers of information from other websites on the Internet (O'Reilly, 2007). A website such as Britannica Online would host an online encyclopedia of knowledge, but its functions would mainly be limited to providing information. After the dot-com bubble, there were websites that allowed users to provide information back to the website in exchange of improved services. The encyclopedia website Wikipedia would foster a rich community of users who would use its website to access its content, but also help in improving its content by co-authoring their articles (Anderson, 2007). The two-way communication of data between users on the internet has allowed websites to become online platforms that are continually updated, combine data from multiple sources, and deliver rich user experiences (O'Reilly, 2007; Tredinnick, 2006). This two-way communication of data can also allow users to share data onto platforms in order to collaborate with each other for various purposes such as writing encyclopedia articles, creating geographic maps, or even computer software (Kittuer & Kraut, 2008; Budhathoki & Haythornthwaite, 2013; Dabbish, Stuart, Tsay, & Herbsleb, 2012).

The term Web 2.0 represents the idea that the Internet network could provide online platforms to all devices (and users) connected to it (O'Reilly, Web 2.0: Compact Definition, 2005). By connecting users on the internet together to online platforms, it is possible for them to create innovations of value by working together towards a unified goal (Baldwin & Von Hippel, 2011). O'Reilly (2007) presented core competencies of Web 2.0 which are outlined in Figure 1. Web 1.0 was referred to websites that mostly provided data to consumers, whereas Web 2.0 was referred to websites that became online platforms who provided two-way communication of data and services between a website and its users. Platforms such as Wikipedia and OpenStreetMap would allow users to act as 'Citizen Sensors' (Goodchild M. F., 2007) and voluntarily contribute data to create a database of knowledge which can then be used for public good, or scientific research. Instead of relying on an authoritative organization to make an encyclopedia or a map, these online platforms would coordinate their users to co-develop their data. When using appropriate coordination techniques, the quality of articles on Wikipedia have been known to improve by adding more editors (Kittuer & Kraut, 2008). Studies have proven that voluntarily generated data on OpenStreetMap meets the 'Linus Law', which is the assumption that the quality of a product increases as the number of contributors increases (Haklay, Basiouka, Antoniou, & Ather, 2010; Haklay M. , 2010). Given that the number of users on the internet has increased from 400 million in the year 2000, to an unprecedented 3.2 billion in 2015 (ICT Facts & Figures, 2015), this creates additional opportunities for collaboration.



Web 2.0

Services, not packaged software, with cost-effective scalability

Control over unique, hard-to-recreate data sources that get richer as more poeple use them

Trusting users as co-developers

Harnessing collective intelligence

Leveraging the long tail through customer self-service

Lightweight user interfaces, development models AND business models

Figure 1: Core competencies of Web 2.0 companies (O'Reilly, 2007)

Researchers have found that there are generally two types of users who voluntarily collaborate on online platforms: (1) serious users who are focused on building community, knowledge, and career, or (2) casual users who are focused on the free availability of the data (Budhathoki & Haythornthwaite, 2013). Utilizing the diverse set of skills, knowledge, and volume of users of the Internet, collaborative innovations can benefit both the public and the private sector (Levine & Prietula, 2014). For example, spatial data generated from OpenStreetMap is used by the private organization such as Mapbox and Carto to provide specialized maps or data visualization services to paying customers. OpenStreetMap's spatial data is used for some government projects that involve mapping of areas that haven't been covered before, because the resulting data could be used for disaster relief or humanitarian aid projects (Haklay & Budhathoki, 2010).

As citizens, private sector, and governments become more digitally connected, online platforms will start to facilitate the interactions between them, however there is little research done to understand how their usage plays into open government initiatives or the collaborations between them. The platforms, users, and innovations made in collaborative settings for use in government work should be further researched to understand its users, collaborations, and resulting outcomes. The following sections explore the definitions and literature surrounding open government, open collaboration, and the online platform GitHub, which is the focus of this research.

1.2 Open Government

'Open Government' is a concept that governments should allow citizens to participate in the decision-making process that affects them, as well as to make government information as transparent as possible so it can be used to create public value (Harrison, et al., 2012). Open government and freedom of information are closely connected, and originate from historic efforts to combat corruption in governments that hide criminal or unethical actions (Taewoo, 2012; Yagoda, 2010). Incidents such as the Watergate scandal in the American government institution has brought the need for governments to become more open about their dealings (Wirtz & Birkmeyer, 2015). Open government initiatives can create public value and improve democracy by sharing their data as 'open data' for the public to use (Kalin, 2014). Since the inception of open government, its definition has been expanded upon to create frameworks of how open government could be understood and evaluated. The following subsections discuss the definition and frameworks of understanding open governments.

1.2.1 President Obama's Memorandum on Transparency and Open Government

President Obama's 2009 Memorandum on Transparency and Open Government discussed three key features for openness in a government that would ensure public trust: transparency, participation, and collaboration (The White House, 2009; McDermott, 2010). Transparency promotes accountability via sharing government information on their actions, policy, and assets, allowing governments to be open to criticism. Participation promotes public engagement into decision making, knowledge sharing, and drawing on the collective knowledge of all stakeholders. Collaboration allows all citizens and governments to cooperate using innovative tools and systems to create innovations of public value (McDermott, 2010; Lee & Kwak, 2012). Although the memorandum was the first major government initiative in the American government to strive for openness, it made a significant impact in order to champion the idea of openness for government across the world (Wirtz & Birkmeyer, 2015).

1.2.2 Open Government Framework

Wirtz & Birkmeyer (2015) defined open government as a framework of transparency, participation, and collaboration in a multilateral, political, and social process between governments and its citizens. These processes are facilitated by modern information and communication technologies that improve the effectiveness and efficiency of governments.

As outlined in Figure 2, they framed transparency, participation, and collaboration as the foundation for governments that can be used to improve society and create public value for citizens.

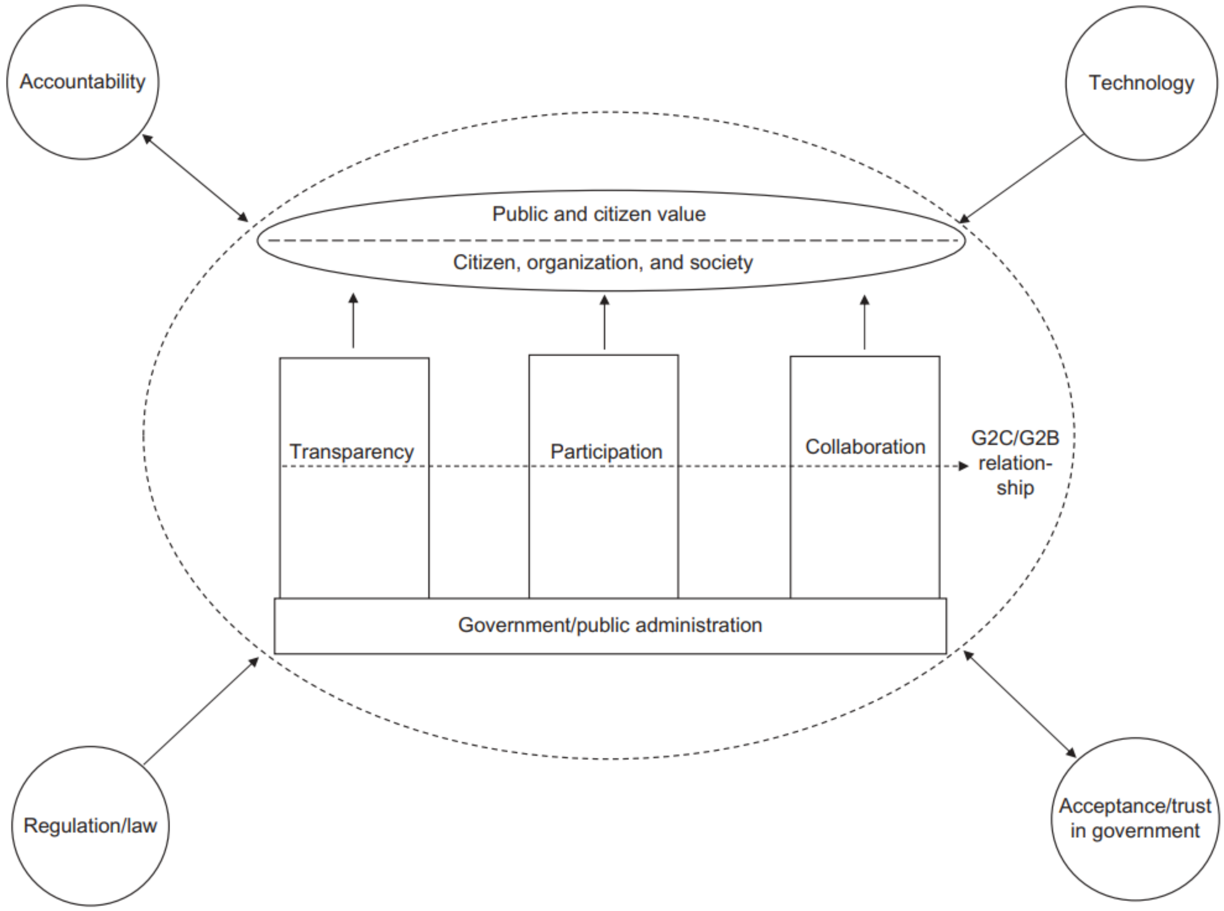


Figure 2: Open Government Framework (Wirtz & Birkmeyer, 2015)

The principles of Wirtz & Birkmeyer’s (2015) open government framework are built upon the definition of Obama’s Memorandum of Open Government, however they add four external factors that impact the effectiveness of an open government which include (1) technology, (2) accountability, (3) regulations, and (4) trust in government (See Figure 2). Technology refers to Web 2.0 technologies that offer interactions between governments and citizens through government websites, social media such as Twitter, or other online platforms. Accountability refers to the idea that governments should be held responsible for their decisions that affect citizens. Transparency of data and decisions can improve accountability of governments and improve public trust and acceptance in governments. Regulations refers to clear and comprehensive legal frameworks within countries, regions, and cultures that govern what governments and citizens can do. Additionally, government-to-citizen or government-to-business (G2C/G2B) relationship refers to strength of the relationship between governments and citizens/businesses that can be improved with increasing levels of transparency, participation, and collaboration. These relationships are influenced by all the external factors of open government where the leadership from government to communicate information, data, and decisions can influence public trust and relationships between citizens and other organizations (Wirtz & Birkmeyer, 2015; Janssen, Charalabidis, & Zuiderwijk, 2012).

1.2.3 Open Government Maturity

Becoming an open government as defined by the Government of Canada's 2020 National Action Plan on Open Government[1] is "an approach to governance that focuses on transparency, accountability, and citizen participation". The plan outlined an approach to enforcing their core principles of inclusion, gender equity, accessibility, user-centric thinking, reconciliation, and collaboration. It proposed various actionable goals or milestones for becoming an open government, however it lacks in proposing methodologies for measuring the performance of open government initiatives. It instead suggests governments to host public surveys, interviews, or open government events in order to gauge the progress of their initiatives.

Although the National Action Plan does not directly address how governments could progress towards becoming an open government, it is possible to reframe the model of open

[1] National Action Plan on Open Government (https://open.canada.ca/en/content/canadas-2018-2020-national-action-plan-open-government)

government proposed by Wirtz & Birkmeyer's (2015) into stages of maturity instead of all the parts being implemented in parallel. Open government progress can be measured by analyzing social media based open collaboration platforms that have been used by governments (Mergel, 2015). Lee & Kwak (2012) proposed an Open Government Maturity Model (OGMM) for social media-based public engagement to assess and guide open government initiatives (see Figure 3). They organized their OGMM into five stages (Level 1) Initial Conditions, (Level 2) Data Transparency, (Level 3) Open Participation, (Level 4) Open Collaboration, and (Level 5) Ubiquitous Engagement. Initial Conditions (Level 1) refers to governments providing basic information about themselves through government websites with one-way communication and little to no public engagement. Data Transparency (Level 2) refers to the making governments transparent by providing government data of high value and quality that is accurate, timely, and conforms to modern data standards. Level 2 allows limited use of social media to gather feedback from the public about the data.

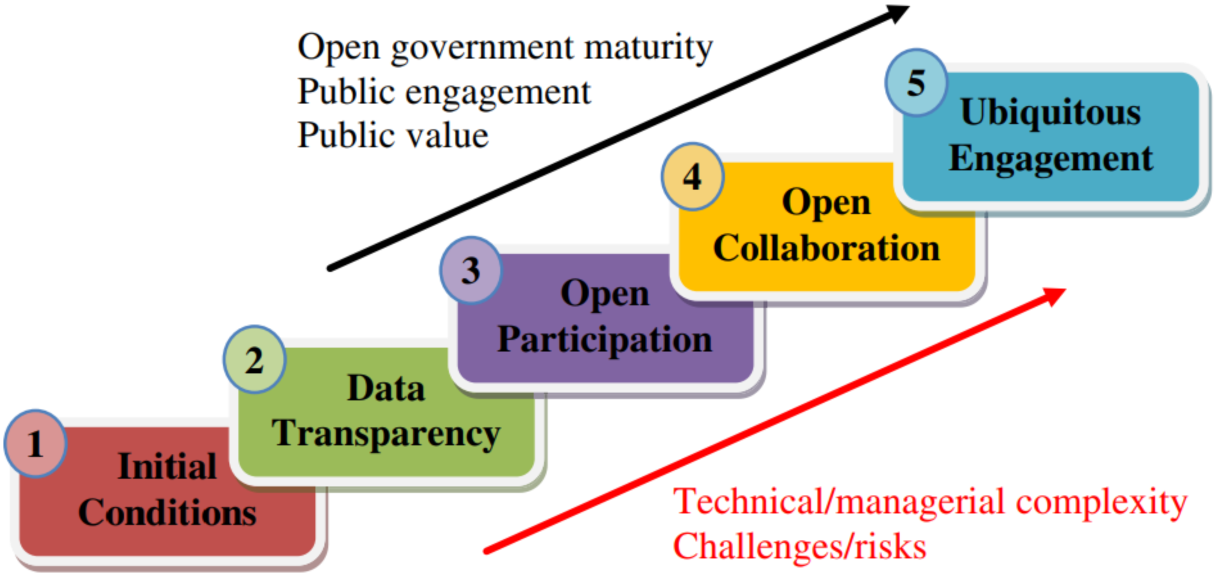


Figure 3: Open Government Maturity Model (Lee & Kwak, 2012)

Open Participation (Level 3) refers to improved communication that allows public feedback and better conversation with the governments. Level 3 allows citizens to participate in voting and ideation processes for government projects with real time engagement and improved sense of community. Open Collaboration (Level 4) refers to collaborations made with other agencies or the public to co-create innovations or services of public value. Level 4 requires Web 2.0 technologies and online platforms to facilitate open collaboration processes for complex projects and decision making. Ubiquitous Engagement (Level 5) refers to an overall improvement to transparency, participation, and collaboration while having seamless communication between agencies and public engagement. Level 5 allows universal access to government data through online platforms, mobile devices, and social media channels. It also focuses on the outcome of open government initiatives rather than the process to operate them in order to focus on the public value of open government initiatives.

Social media based online platforms can be used by governments to reach a broader group of citizens to take part in their initiatives. For example, some governments have used online platforms in order to facilitate participatory public policy making, or forecasting political opinions, or even measuring noise pollutions using smartphones (Yannis & Euripidis, 2012; Sobkowicz, Kaschesky, & Bouchard, 2012; Maisonneuve, Stevens, & Ochab, 2010). Existing Web 2.0 platforms that offer social media like communication and participation between citizens can help governments skip the development costs to create these communication tools and jump ahead in their levels of open government maturity levels. Utilizing already existing social platforms that offer government organizations to become open has become a major trend in electronic government (e-government) practices worldwide (Criado, Sandoval-Almazan, & Gil-Garcia, 2013). There is a need to research how these governments are using these platforms and how it affects their progress towards becoming an open government.

1.3 Open Collaboration

'Open Collaboration' is a concept that anyone can voluntarily work in a project for any reason, and the results of the project can be shared with everyone without restricting user to modify, or repurpose the results of the collaboration (Baldwin & Von Hippel, 2011). Web 2.0 technologies enable open collaboration because they allow online platforms to facilitate open collaboration between a distributed set of contributors with varying levels of usage and expertise (Goodchild M. F., 2007). Academic literature defines open collaboration as both a process and a system (Forte & Lampe, 2013; Levine & Prietula, 2014). The following subsections discuss these competing definitions of open collaboration and give examples of online platforms that use it in practice.

1.3.1 Open Collaboration Frameworks

Forte & Lampe (2013) define open collaboration as distributed, collaborative efforts made possible because of online technologies that facilitate collaborative activities. They outline open collaboration as a set of common characteristics that (1) support the collective production of an artifact, (2) technologically mediate the collaboration platform, (3) have a low barrier to entry and exit, and (4) support the emergence of persistent but malleable social structures (Forte & Lampe, 2013). Collective action represents the idea an action can be taken to benefit an entire group rather than one or a few members (Van Zomeren & Iyer, 2009; Wright, Taylor, & Moghaddam, 1990). Collective production (or action) is a core part of open collaboration because it requires multiple persons or organizations to collaborate with a shared set of goals to create data or Intellectual Property (IP) (Budhathoki & Haythornthwaite, 2013). Collective action can be taken by a community of people for political reasons such as enacting against gender or racial discrimination (Morris, 1986; Kelly & Breinlinger, 1996), or for altruistic reasons to create innovations of public value (Baldwin & Von Hippel, 2011; Baytiyeh & Pfaffman, 2010). Collaboration platforms can not only mediate the collaboration between users, but they can also provide the means to develop social structures and operate as a social network between users. Records of social interactions can be stored and used to generate social capital among the users, i.e. users subscribing or following other users (Resnick, 2001). The process of open collaboration is focused around an idea which brings together a community of people who want to collaborate, with their actions being facilitated by a unifying platform.

Levine & Prietula (2014) define open collaboration as a system of distributed users on the Internet that contribute their work to create innovations (product or data) of economic value that are shared publicly to all collaborators and non-collaborators alike. They also outline elements of open collaboration which are to (1) create goods of economic value, (2) allow open access in contribution and consumption of data, (3) make interactions central to the system, and (4) allow participants labor to be purposeful yet loosely coordinated (Levine & Prietula, 2014). A product made through open collaboration has 'economic value' if it can be substituted with a for-profit version of the same product (Edelmann, Höchtl, & Sachs, 2010; Quelin, Kiveleniece, & Lazzarini, 2017). In this context, Linux, the free and open-source computer operating system made through open collaboration would have economic value because it can be substituted for commercially purchasable computer operating systems such as Microsoft Windows. 'Open access' implies that the distribution of the IP generated from the open collaboration systems do not discriminate in providing access to any persons, groups, or other technology (Open Source Initiative, 2019). Interactions are central to coordinating open collaboration but are not as formal as traditional organizational hierarchies, i.e. employer-employee structures. In an open collaboration system, users are free to self-organize by defining goals and responsibilities that they can volunteer to complete (Steinmacher, Conte, Gerosa, & Redmiles, 2015). Overall, the collaboration system can facilitate open collaboration, but with a focus on creating innovations of economic value.

Levine & Prietula (2014) defined open collaboration from an organizational perspective, where the contributors are a means of achieving the goal of an open collaboration project. Comparably, Forte & Lampe (2013) defined open collaboration from a collaborator’s perspective where the process of collaboration and social interactions among the users are just as important as the final product. Instead of considering open collaboration as a system to generate goods of economic value or social networks, it could be synthesized more broadly as an approach to create IP through the collaborative contributions of a distributed set of users who are all connected and mediated by an online central platform. By considering open collaboration as an approach rather than a system, it reduces the importance of the platform facilitating the collaboration and instead focuses on the shared goal of creating the IP.

Utilizing an open collaboration approach brings both opportunities and challenges. Open collaboration systems need to have a low barrier to entry and exit where user can contribute as little or as much as they want (Schneider, 2013). With platforms such as Wikipedia or OpenStreetMap, for a user to enter the platform and start collaborating, all that is required is to access the website via the Internet, and to create an account. As for exiting the platform, users can quit anytime, with no employer or contract to obligate them to continue their collaboration. A low barrier to entry also increases the number collaborators to a platform, however the skill, knowledge, motivations, and amount of contributions that users provide will vary (Lampe, Wash, Velasquez, & Ozkaya, 2010). Due to the amount of participation not being equal, there are concerns of open collaboration projects being of lower quality and completeness (Haklay M. , 2010). However, as the number of collaborators increases, the overall quality of a product can increase because although one person may not be skilled in every aspect of the project, many individuals can work together to combine their skills and compensate for the lack of knowledge from any one person (Kittuer & Kraut, 2008; Budhathoki & Haythornthwaite, 2013). Studies have proven that the data generated in open collaboration approach on OpenStreetMap meets the Linus Law and it’s quality should keep getting better over time as the number of creators increases (Haklay, Basiouka, Antoniou, & Ather, 2010; Raymond, 1999).

1.3.2 Open Collaboration in Practice

There are various examples of platforms that each focus on certain shared goals that are being met by using an open collaboration approach. Wikipedia, an encyclopedia website that is made through the contributions of hundreds of thousands of users that have collaboratively written over 49 million articles in 250 languages, is the most common example of an open collaboration platform, (Wikipedia Statistics: All Languages, 2019). This type of collaboration is an example of 'crowdsourcing', where an action formally performed by a member within an organization can be outsourced to a distributed network of individual volunteers through web-based solutions (Howe, 2006; Brabham D. C., 2008; Brabham D. C., 2010). Although an individual may be limited by their own individual skills, by having a large group of individuals committed to solving a problem, the collective knowledge or the 'wisdom of the crowd' can be utilized in order to solve the problem by using the best ideas from a group of individuals (Kittuer & Kraut, 2008).

Crowdsourcing has been used to gather accurate information about a topic through the input of a large volume of citizen sensors (Goodchild & Glennon, 2010). Wikipedia has proven itself to be of better quality than other for-profit encyclopedias due to its larger volume of contributors (Giles, 2005). Any issues in the data can be rectified at a much greater pace than traditional for-profit encyclopedias due to crowdsourcing for information instead of only relying on a closed set of experts or lengthy update schedules. The open collaboration approach of Wikipedia can also bring some issues because it allows its content editors to stay anonymous (Santana & Wood, 2009). Although this may protect the identity of the user, the lack of transparency of who is editing the content and for what purposes can bring about social or ethical consequences. Additionally, there are cross-cultural and political issues with non-English versions of Wikipedia where its content is driven towards the voice of the political majority (Liao, 2009; Santos & Cabral, 2009). For example, the Chinese version of Wikipedia has over four dialects that caters to different regions of Asia, however the translations of their content between one dialect to another shows variations in the context of the content from the different political regions such as Hong Kong, Taiwan, and China.

Geographic information collected in a crowdsourced approach is often referred to as Volunteered Geographic Information (VGI) (Goodchild & Glennon, 2010). OpenStreetMap is an example of an open collaboration platform where users provide VGI about real-world geographic features to create a freely available online map. OpenStreetMap has over 5.9 million users accounts on their platform that has shared 7.6 billion GPS points, and created over 6.3 billion points of interest, roads, regions and other geographic data (OpenStreetMap Stats Report, 2019). Although the platform has many registered users, only about 1% of its registered user accounts are active each month, with around 250,000 active contributors per year (OpenStreetMap Stats Report, 2019). The data generated from Wikipedia and OpenStreetMap are shared as 'open source' products, which represent IP that is released under a public license allowing anyone to use, modify, or redistribute the IP without discrimination to persons, groups, fields, or other technology (Open Source Initiative, 2019). The open collaboration approach of OpenStreetMap also has some issues in how its data is generated which stems from the influence of geography, and participation from its users. OpenStreetMap collects geographic data about real world assets, and it provides an extensive coverage of aerial imagery to speed up the mapping of urban landscapes. Although users map based on the aerial imagery available, for areas too difficult reach that can only be mapped by ground surveys will have fewer users willing to contribute geographic information there because users are either unwilling or unable to go there. Additionally, users will mostly contribute information for areas they are most comfortable with, leaving large areas unmapped even if it has aerial imagery available because nobody lives or goes in those areas or is knowledgeable enough to map them (Haklay & Weber, 2008).

1.4 Open Collaboration on GitHub

Open collaboration is also used to create open source software where users contribute changes to a repository of code that is shared publicly on an online platform such as GitHub, Gitlab, or BitBucket. GitHub is an example of a platform that allows users to create computer software in an open collaboration approach, and it is used worldwide for open source (public) and private software projects (Peterson, 2013). It is built upon the Source

Code Management (SCM) tool called Git, which is a robust framework designed to allow for any number of users to contribute changes to any repository of information which could be in the form of computer software code, datasets, text documents, and more (Dabbish, Stuart, Tsay, & Herbsleb, 2012; Peterson, 2013; Storey, Singer, Cleary, Figueira Filho, & Zagalsky, 2014; Peterson, 2013). Although GitHub relies upon Git to handle its source code, it acts as a version control system that distributes and coordinates the collaboration between all users on its platform.

Since its creation in 2007, GitHub has been gaining popularity for hosting open source projects, which also attracts governments who want to use the platform to host their own work, or use open-source software for their own projects (Longo & Kelley, 2016). GitHub has over 23 million user accounts working on over 18 million public repositories of data (GitHub Repositories, 2019; GitHub Users, 2019). Due to having relatively no cost to acquire open source IP, and the approach allowing for a large volume of collaborators, governments have been shifting towards using open collaboration on GitHub for some of their own work (Longo & Kelley, 2016). The enforcement of collaboration and transparency of information makes it the ideal platform for governments to adopt in order to comply with their own mandates of open government, specifically to satisfy their e-government aspects when it comes to creation or procurement of computer software.

1.4.1 Coordinating Collaborations using Git

GitHub relies upon Git to support the collaboration between its users. Git is a version control system that is a freely available open source project that was created by Linus Torvalds in 2005 to help develop the Linux kernel which is an open source operating system. As illustrated by Figure 4, Git allows users to manage their source code by handling all versions of the code that exist through three major operations which are (1) fork code, (2) create pull requests, and (3) merge changes. A repository of code can be created by any user, and different version of the repository are managed as different branches.

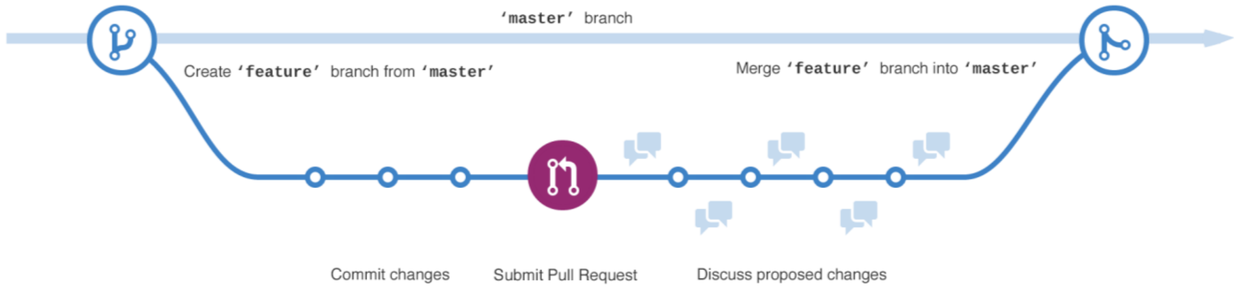

'master' branch
Create 'feature' branch from 'master'
Merge 'feature' branch into 'master'
Commit changes
Submit Pull Request
Discuss proposed changes

Figure 4: Version control system workflow (GitHub Guides, 2019)

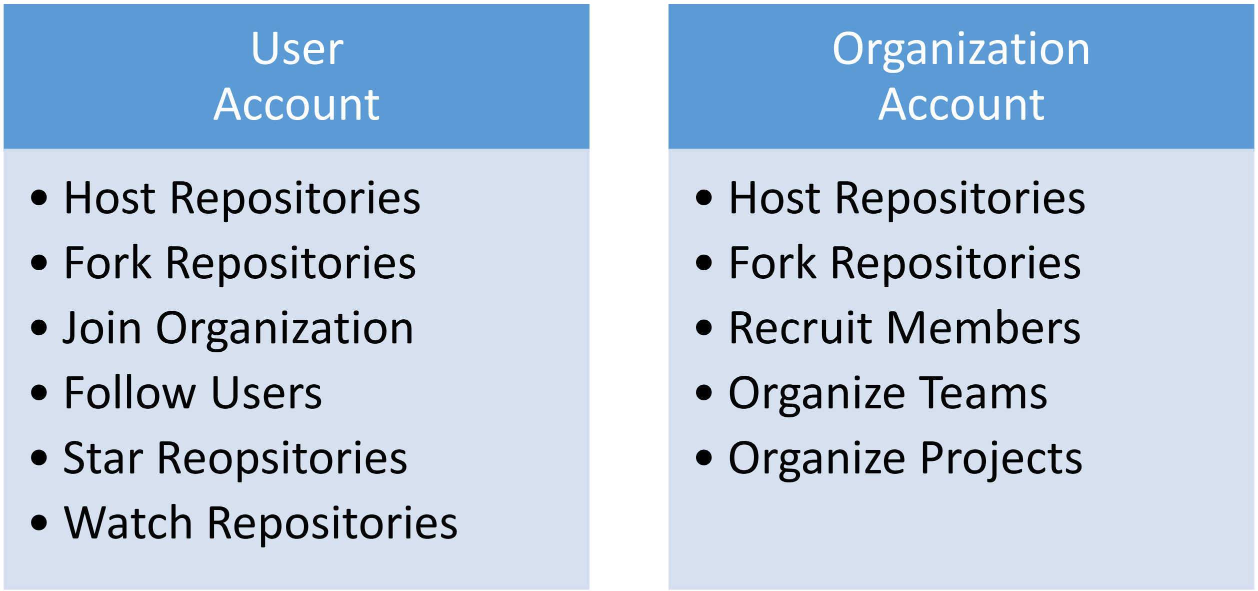
The main version of the repository is referred to as the 'master' branch. Users 'fork' code where they create their own branch or version of the master branch called a 'feature' branch in which they can make any edits necessary. These edits can be in the form of inserting or deleting lines of code. Modifying lines of code is handled as inserting new lines with the modified content, while deleting the old versions of the same line. Any edit in the repository must be officially logged by Git as a 'commit' where users can make any number of commits to their feature branch. After a user makes their desired changes, they can request for their code to be merged into the master branch by creating a 'pull request'. Other users can vet the code submitted in the pull request for any errors or issues. Once the code in the feature branch is considered appropriate to be merged into the master branch, the pull request can be accepted. Users can keep working on their feature branch even after the pull request is accepted, or they can close the feature branch and make another one for new features (GitHub Guides, 2019).

Any change to any file in any branch of the repository is logged by Git. Two of the major services that GitHub provides is to (1) act as an online storage by making a copy of the log of these changes, and all the version of the repository that exist, and (2) to re-distribute the data back to all users of GitHub who want to collaborate on the work of any repository. This allows GitHub to act as a mediator to coordinate all their interactions (GitHub Guides, 2019). Although there are other version control platforms that offer similar services such as

GitLabs, BitBucket, Beanstalk, or AWS CodeCommit, GitHub was one of the early adopters of Git and open collaboration which had made it very favorable for open source software projects.

1.4.2 Social Coding on GitHub

GitHub not only acts as an online storage for the project repository, but it also has features that make it into a social platform for its users. As illustrated by Figure 5, there are two types of accounts that can be created on GitHub which have some unique or overlapping features. 'User' accounts are the default account type which represent any individual who has made an account on the platform, whereas an 'Organization' account is an upgraded version of a user account with project management features to coordinate users, projects, and control repositories. Both the users and organization accounts can host repositories themselves, or they can fork a repository from another account. Repositories hosted by any account type can be kept private or listed publicly on GitHub for any other account to see. User accounts can follow other users, allowing them to keep up with their activity on the platform. Users can also Star (favorite) repositories they like, or Watch (subscribe) them to keep up more closely to any updates from them.



User Account

- Host Repositories
- Fork Repositories
- Join Organization
- Follow Users
- Star Reopsitories
- Watch Repositories

Organization Account

- Host Repositories
- Fork Repositories
- Recruit Members
- Organize Teams
- Organize Projects

Figure 5: GitHub account types and features

Users can join organizations via invitation in order to access repositories that may be kept private or have special restrictions on who can contribute to them. By default, users who have joined an organization are privately listed as a member of the organization, and they would need to voluntarily allow themselves to be publicly listed if they want others on GitHub (outside the organization) to see them as a member of it. Organization accounts on the other hand can also host repositories of their own, but they are outfitted with more project management features allowing them recruit members and organize them into teams and assign project tasks. Users inside an organization can be assigned specific privileges as to what repositories they have access to or join teams with specific privileges. Organizations can also coordinate development initiatives by having projects that allow them to organize tasks, roles, and responsibilities for users to complete.

Analyzing social coding through GitHub is a field of research that aims to study open collaboration between users, and the outcomes of their interactions. Prior literature in the field collects data from GitHub in order to analyze its uses for software education (Zagalsky, Feliciano, Storey, Zhao, & Wang, 2015), developer communication (Storey, Singer, Cleary, Figueira Filho, & Zagalsky, 2014), transparency in coding (Dabbish, Stuart, Tsay, & Herbsleb, 2012), employability (Rusk & Coady, 2014), diversity (Vasilescu, et al., 2015), and more. Due to the nature of GitHub being friendly to open source projects, governments have been adopting the platform to host their own projects, collaborate in the open, and overall become an open government. Although research into open collaboration, and open government exists in their own separate literature, little research has been done to analyze them both via GitHub. Mergel (2015), analyzed open collaboration in the public sector in the United States by using GitHub data to understand which organizations are collaborating through reusing code (forking code) or contributing to other organization's code (sending pull requests). Longo & Kelley (2016) analyzed GitHub use in the public sector in Canada by directly interviewing government workers who have used GitHub for collaboration in their work. However, there is an opportunity to combine prior research techniques of analyzing open collaboration and open government to further study them together through government use of GitHub.

1.5 Research Goal

The overall goal of this research was to develop a better understanding of why government organizations use GitHub for collaboration, both internally, and with external contributors. The research was based on the following two questions below:

Research Question #1: what are the strengths, weaknesses, opportunities, and threats faced by Canadian governments collaborating on GitHub?

Research Question #2: what is the extent and nature of government collaboration on GitHub? Specific areas of focus for the research question listed below:

1. Are government organizations actively using GitHub? If so, how many?
2. How complete is the information available about government organizations on GitHub?

3. Is there a collaborative relationship between government organizations that use GitHub?

1.6 Research Scope

GitHub contains almost 22 million user accounts and 3 million organization accounts. Although anyone can create an account on GitHub and start collaborating on computer software, only a small proportion of organizations are government accounts with repositories dedicated for government or public use. Creating a framework for searching through all of GitHub for actual government accounts would require vast amounts of time and computational resources and was beyond the scope of this thesis. GitHub itself keeps a track of which organization are actual government accounts via a crowdsourced list called the Government GitHub Community (GGC) (https://government.github.com/community/) list, and it will be used instead of manually searching through all of GitHub. The GGC list contains the names of over 782 organizations that originate from 59 different countries which control 31,703 repositories of data, to which 71,549 users have collaborated on via making commits. The scope of research will be limited to only include the organizations, repositories, and users in the GGC list. Data collected about these organizations in this research is up to date as of July 2019.

1.7 Thesis Outline

This thesis is organized around its two research questions. The first question was to explore the motivations for Canadian governments to use GitHub, while the second question was to explore the extent and nature of government collaborations on GitHub. In the first question, Canadian government users of GitHub were interviewed using open-ended questions about their usage of GitHub. Their perspectives were organized into a SWOT analysis with the following categories: Strengths, Weaknesses, Opportunities, and Threats

(Pickton & Wright, 1998). Themes uncovered from the SWOT analysis are used to determine the motivations and challenges governments may face when using GitHub.

In the second question, data was collected about government organizations in the GGC list using GitHub's Application Programming Interface (API). The collected data was explored and analyzed to determine the types of governments, their activity, and the quality of information available about them. A social network graph of government organizations was created, analyzed, and mapped geographically to examine who they are collaborating with. Analyzing the data from GitHub would reveal which the scope and extent of the collaborative nature of these government organizations, and to see which are the most collaborative and why.

The first research question was discussed in Chapter 2, while the second question was discussed in Chapter 3. Lastly, Chapter 4 brought together the findings from both the chapters and contextualized them with reference to existing academic literature. Chapter 4 provided conclusions on governments' adoption of online platforms for collaboration, as well as presented directions for future research.

Chapter 2

2 Motivations and Challenges to Collaboration: A SWOT Analysis of Government use of GitHub

2.1 Introduction

The use of GitHub within Canadian governments and its public sector is relatively new when compared to its use in the United States of America (USA) (Longo & Kelley, 2016; Lima, Rossi, & Musolesi, 2014; Mergel, 2015). GitHub contains 49 Canadian government organization accounts with 1356 public repositories, and 246 public members. Although these numbers may be small compared to USA which has 300 government organization accounts with 10,207 repositories, and 1150 public members, there is still a significant amount usage of GitHub in Canada. Prior studies have tried to analyze the social coding side of GitHub by surveying or interviewing some of the users of the platform (Dabbish, Stuart, Tsay, & Herbsleb, 2012; Zagalsky, Feliciano, Storey, Zhao, & Wang, 2015; McDonald & Goggins, 2013), however very few studies have analyzed its use for public sector or government work (Mergel, 2015; Longo & Kelley, 2016). The following subsections reviews previous literature on analyzing government use of GitHub and proposes the research goal of this chapter.

2.1.1 Open Collaboration via Pull Requests

Mergel (2015) analyzed forks and pull requests made by GitHub organizations from the USA in order to understand how they were sharing code with each other, as well as interviewing government organization managers about their usage of the platform. An organization forking another's repository was considered as them reusing their code. Adding features to the fork and making pull requests for them to be merged with the original repository was considered as engaging in open collaborating. Mergel discovered that although the network of forking collaborations could not explain why public workers were forking repositories or if they could make use of it, forking in general was more common than submitting pull requests. This was partly due to the ease of access when forking repositories because it was a one-click action to make forks, rather than the process of making pull requests that required the time to make sure the new changes were compatible with the existing code.

The reuse of code had become an important mechanism in governments which allowed them to be more transparent, engage the public, and open silos of knowledge. Although not all the code from a project was inherently useful just by making it public, some managers discovered that providing small packages of code that served specific purposes would be more useful to themselves and others on the platform rather than the code from larger projects which not everyone would be able to use or know how to use. Open sourcing code and coding in the open gave others the opportunity to participate by reviewing the code or making pull requests. Since GitHub allows users to comment on code in a line by line basis, any user of the platform can provide granular feedback and engage in thorough discussions about the contributions in the repository. Although it was helpful for organizations to receive pull requests from other developers, not all pull requests were accepted. Every pull request would undergo through review of the contributed code, as well as evaluating the reputation of the developer. The quality of the code and the direction of the project would be prioritized in order to avoid bad code. Code from inexperienced developers would be required to be rewritten until it was up to the quality of the organization. Developers unable to or unwilling to improve the code would have their pull requests rejected. Overall, reusing and adapting code from open source repositories that allow for open collaboration through GitHub had been successful to develop innovations for governments in the USA, however there was a lack of understanding on how the platform performs with other countries, and how it would help them become more open (Mergel, 2015).

2.1.2 Open Collaboration for Transparency

Longo & Kelly (2016) analyzed the use of GitHub within Canadian public service as an early look into the tool for open collaboration within Canadian governments. They conducted surveys and interviews with users from GitHub that were members of Canadian government organizations and found that 53% of the user showed no activity between June 2014 and June 2015. Of the active users, there was a long-tail distribution of contributions where a small number of users were doing most of the work while the majority were rarely active. They also discovered that although half of the respondents considered themselves to be fluent in the use of Git and GitHub, 63% claimed to have little to no influence on getting their workplace to adopt the use of GitHub. Of the participating organizations, open collaboration was a way to (1) become efficient for coordinating projects, (2) allow different departments to communicate information and resources, and (3) to create innovations for users inside and outside the government (Longo & Kelley, 2016). GitHub allowed government workers to overcome bureaucratic constraints of procuring software, as well as respond to changing requirements.

It also offered some advantages from open sourcing software because it was a useful tool to respond to freedom of information requests, as well as providing information to the public in both of Canada's official languages (English and French). For example, the Canadian Charter of Rights and Freedoms ruled that government websites needed to be available in both of Canada's official languages, and in response the Treasury Board of Secretariat (TBS) Canada helped create the Web Experience Toolkit (WET) as an open source web framework on GitHub to serve that demand, as well as develop it openly on GitHub with the input of other government organizations within Canada. The use of GitHub was generally accepted by governments, but its implementation was limited by the lack of technical knowhow to use the platform, as well as the dependence on other departments to collaborate on time sensitive projects. Although GitHub could be dismissed as another communication tool, in theory it provided a new approach to collaboration that could have profound improvements on how governments function within themselves, with other departments, and with the public.

2.1.3 Open Collaboration to Advance Open Government Policies

GitHub can allow governments to become more open because it was designed for open collaboration and to make open source software in a social coding environment. Users accounts on GitHub can be from the general public, public sector workers, contractors, or 'civic hackers' who want to improve governments through data activism, advocacy for transparency, and political participation (Schrock, 2016). Organization accounts are from federal, regional, and local level, as well as other for profit and not-for-profit (NFP) organizations that want to build civic tech. Re-framing the progression of becoming an open government into the Open Government Maturity Model (OGMM) proposed by Lee & Kawk (2012), GitHub can help governments fast track their progression to become an open government in many ways. Governments can achieve Level 1 (Initial Conditions) by simply having an account on the platform with information about their government organization. Level 2 (Data Transparency) can be achieved by having their projects be publicly available on GitHub for others to reuse or fork. Level 3 (Open Participation) can be in the form of allowing users to participate in the deciding the progression and priorities of any project via reviewing the code and discussing issues. Level 4 (Open Collaboration) would be in the form of accepting pull requests from users on the platform who aren't just government workers, but people willing to help write code for a project.

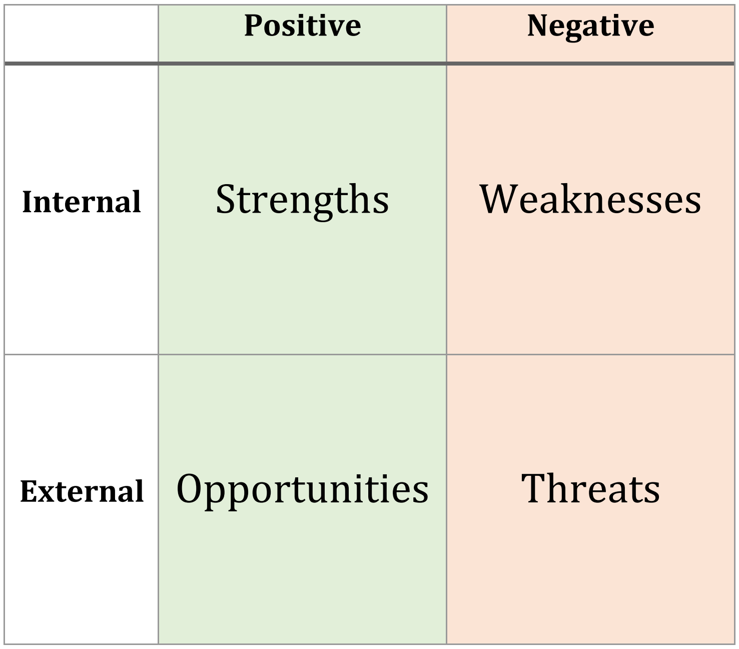
2.1.4 Research Goal

Although GitHub can allow organizations to become an open government as it has been for many organizations in the USA, its use within Canada is still growing and it is still uncertain if it will become a mainstream approach for Canadian governments to collaborate on. Thus, the research goal of this chapter was to understand the motivations of why current

Canadian governments use the platform and to understand the challenges they faced in order to use it. Completing the research goal will help government workers, policy makers, and the public make better decisions about what they want the future of collaboration for civic tech in Canada to be.

2.2 Methods

In order to complete the research goal, Canadian government workers who have used GitHub for government related work were interviewed to get their perspectives about the platform. Themes identified from their interviews were organized into a SWOT analysis consisting of the following categories: Strengths, Weaknesses, Opportunities, and Threats. SWOT analysis has been traditionally used in business analysis to evaluate the features of a product or service and how it affects the user. It is a framework used to analytically categorize significant environmental factors internal and external to an organization (Pickton & Wright, 1998). In this case, GitHub was evaluated for its use to Canadian governments. As illustrated by Figure 6, Strengths and Opportunities refer to positive characteristics of the product, whereas Weaknesses and Threats refers to negative ones. Strengths and Weakness are about factors internal to the product, whereas Opportunities and Threats refers to external factors that affect the use of the product. Organizing the findings from the perspectives of the interviewees into the SWOT structure helped uncover themes about the usage of GitHub for government work.



	Positive	Negative
Internal	Strengths	Weaknesses
External	Opportunities	Threats

Figure 6: SWOT analysis categories

All themes listed in their respective categories of the SWOT analysis chart were sorted in their number of occurrences from the perspectives of the interview participants. Additionally, each bullet point representing a theme will also contain a tally of the number of participants that discussed a theme in order to quantify the significance of the theme or give it some weight in contrast to other themes. Each theme was developed by carefully reading through the interview transcripts and notes and identifying repeating idea, quotes, and discussion that could fall under a theme. Although the themes were grouped into distinct categories of the SWOT analysis, a few of the themes represented a mix of ideas which were fluid or could overlap into multiple categories. In such cases, the idea was placed in a SWOT category that best represented it by having most of the participants discuss the idea in a way that could be represented as that theme.

Interview candidates were selected from all users accounts on GitHub that were listed as public members of a Canadian government organization from the GGC list. Although there were 246 user accounts that were publicly listed as members of a Canadian government organizations, only a few had an email address associated with their user account.

Additionally, an open invitation to participate in the interview was posted to Government of Canada Collab (GCcollab.ca), which is communication portal like Facebook, but it is designed for Government of Canada workers and Canadian citizens. 33 emails (combination of users and organization contact email addresses) were invited for interviews, from which eight users participated. Full list of interview participants is listed in Appendix A. All interviews were conducted over telephone calls which lasted between 30 to 50 minutes. All interviews were recorded, transcribed, and manually analyzed by going over the transcript to identify outstanding or overlapping themes.

Participants were asked questions about why they use GitHub, what are the advantages and disadvantages they experienced from its use, for example 'What are the benefits you have experienced through the use of GitHub?' or 'What are some of the drawbacks or issues you have experienced through the use of GitHub?' (Appendix B Question 3 & Question 4). Additionally, they were asked if their organization allows open collaboration through contributions from users outside their organization, and if using the platform has improved the transparency of their organization (Appendix B Question 7 & Question 8). The overall goal of the interviews was to directly gather the perspectives from the current Canadian government users of GitHub, and to find motivations or challenges they have experienced from its use that could help other organizations who are considering the use of the platform. Although the interview questions didn't explicitly ask them about their motivations and challenges, the questions were designed to be open ended and gather a broader perspective on their usage of GitHub. The questions were kept broad enough to engage discussion about the long-term use of GitHub, their perspectives about open collaboration, and if the platform allows them to become an open government.

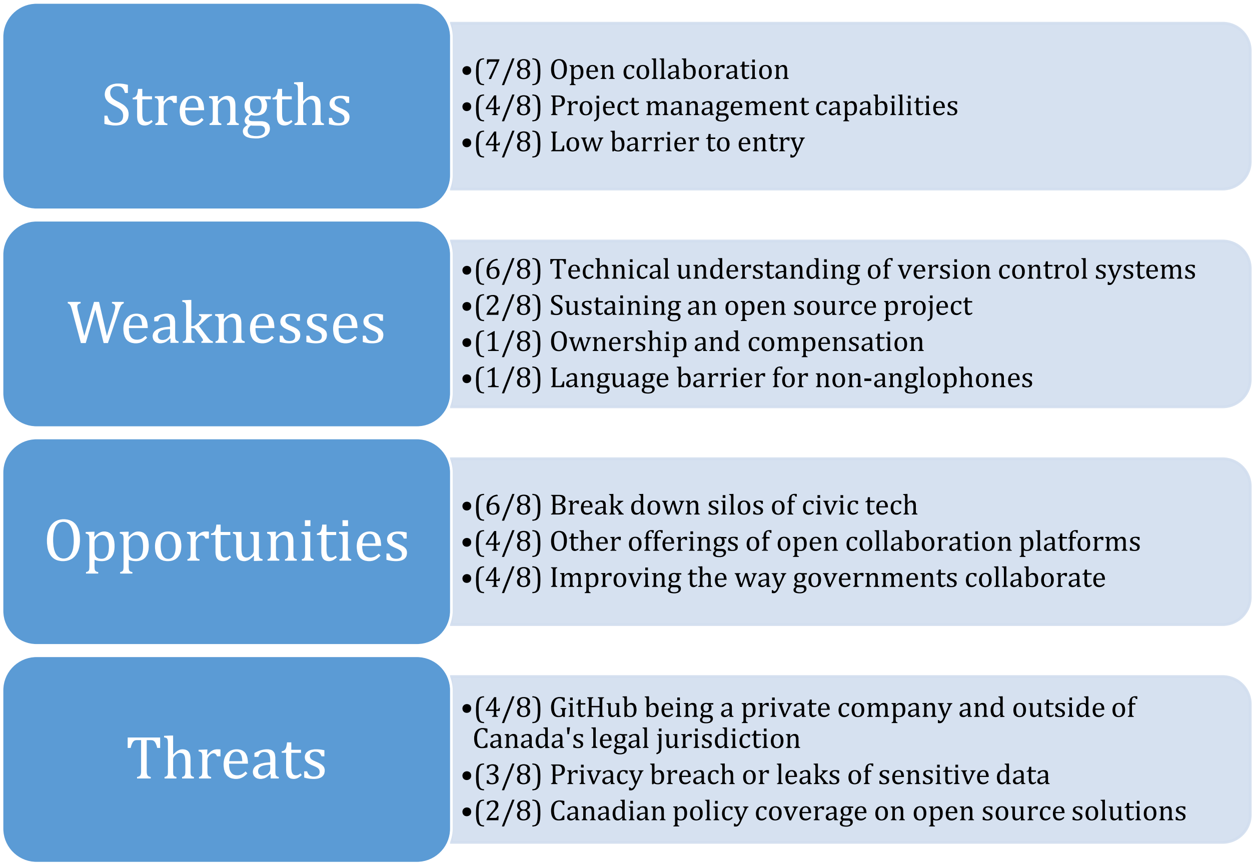
2.3 Results

The Canadian government users who did participate in the interviews were mostly from Eastern Ontario, specifically from Canada's capital city Ottawa, and from the City of

Toronto. There were five participants from the federal Government of Canada offices which included Treasury Board of Canada Secretariat, Privy Council Office, and Shared Services Canada. There was one participant from Environment and Climate Change Canada, and two from the Government of Ontario. Although there are other Canadian government workers in other parts of Canada who use GitHub, it was not possible to contact them because they were either working anonymously on GitHub for their government department, or they choose to not include any contact information in their public profiles, or they simply weren't interested to participate.

There were roughly two types of interview participants: (1) ones who did mostly managerial work, and (2) ones who were coding or developing software. The type of work they did is listed in Appendix A, but the exact job title they had has been omitted for maintaining the privacy of the interview participants. Each type of participant provided perspectives from their own point of view to the interview questions by focusing on the themes that relate to their own work. The managerial workers provided more feedback about project management capabilities, ease of understanding of the platform, and the policies related to their use of the platform. The software developer participants provided more feedback about the open collaboration nature of the platform, the easy of entry and exit, data privacy, and how open collaboration could change the way governments communicate information.

Of the eight participants from the interviews, none of them worked in the same department, or shared the same technological skills or restrictions, however they all had overlapping experiences with their use of GitHub that are summarized in the SWOT analysis in Figure 7. The following subsections will expand on the Strengths, Weaknesses, Opportunities, and Threats that were experienced by the interview participants. The results sections follow the order of the SWOT categories and its bullet point in Figure 7.



Strengths

- (7/8) Open collaboration
- (4/8) Project management capabilities
- (4/8) Low barrier to entry

Weaknesses

- (6/8) Technical understanding of version control systems
- (2/8) Sustaining an open source project
- (1/8) Ownership and compensation
- (1/8) Language barrier for non-anglophones

Opportunities

- (6/8) Break down silos of civic tech
- (4/8) Other offerings of open collaboration platforms
- (4/8) Improving the way governments collaborate

Threats

- (4/8) GitHub being a private company and outside of Canada's legal jurisdiction
- (3/8) Privacy breach or leaks of sensitive data
- (2/8) Canadian policy coverage on open source solutions

Figure 7: SWOT analysis summary of the themes gathered from Canadian government workers who have used GitHub. The fractions corresponding with each bullet point represent the number of participants who had mentioned ideas relating to that theme

2.3.1 Strengths

The most notable strength that participants commented on the use of GitHub was how the platform encouraged open collaboration. Participant F said that “the ability to collaborate […] is the primary motivating reason (to use the platform). The other part is the version control. […] The ability to manage versions as we need to, to be able to go back”. Any change that is contributed to a repository can be uploaded onto GitHub’s servers, allowing it to act as a backup or redundancy for a project’s data. This information can then be delivered back to all users on the platform in multiple ways, most commonly to display the repository on GitHub’s website, or to allow users to clone the repository and make their own changes.

This is effective when working with "large datasets (or projects) and everyone has different directions to exploring those datasets, the ability to collaborate both on analysis, results, and on model development is pretty huge part of it" (Participant F). GitHub allowed government users to collaborate on any project while maintaining control over the IP, and who gets to contribute changes to it.

In order to keep track of all these changes and versions of the repository, GitHub relies heavily upon Git, the underlying version control system which allows a repository of information to be edited by multiple users at any given point, and providing a mechanism to merge all the changes into one version of the repository. By giving user of GitHub the access to any organization's project (assuming the user has permission to access the repository), GitHub can function as a middleman to coordinate the software development for a team of developers. There are instances when "one group wants to view the entire project, maybe one group has their own little changes and they don't want to share those changes because it is their own weird personal change. There are a lot of different rules when you want to put together a project like that […] so Git will do all of things very well" (Participant B). By displaying the project repository in its entirety, all stakeholders are involved in the project, and not just the software developers with technical expertise. If a project is made publicly available, then any user of GitHub can view the project, keep track of any changes that happen, and clone its code to try it for themselves, or even contribute their own changes back to the project repository. Users who are working in diverse teams involving software developer and non-developers are realizing "for collaboration purposes, maintaining computer code, or even to use that code for other actions, (version control) is really important as we see the line between programmer and everybody else shrinking" (Participant F). As more people with diverse skillsets are needed on a project, collaboration software such as GitHub will become a standard that all team members will need to know how to use.

When asked about when GitHub was successfully used for a government project, some of the participants responded with a few examples that have been success stories for organizations which use the open collaboration nature of GitHub to their fullest. The province of British Columbia has a GitHub organization called BC Developer's Exchange[2] where they created a hub of projects that are hosted by the provincial government, but the software was created in collaboration with developers within the government, contract workers, and the general public. The project was successful enough that it encouraged the Government of Canada to create their own Developer Exchange hub[3]. Another example of an open collaboration project was the Web Experience Toolkit (WET)[4] which was designed as an open source library to make government websites more user friendly to the Canadian users of online government services by supporting mobile screens, screen readers, and multilingual users. The project has over 60 contributors, 1000 stars, 600 forks, and still gets regular updates.

It isn't just software projects, but policy documents have been created by government workers as well. On GitHub, the Treasury Board of Canada Secretariat (TBCS) drafted a document called Open First Whitepaper[5] which was designed to encourage governments to shift to an open approach for government initiatives. The document outlines open standards, software, and culture that can be followed to make any government organization progress towards becoming an open government. The group even created their own Open Government License[6] to protect open source projects made by users within Canada. The TBCS also made an Open Source Advisory Council[7], a Digital Playbook[8], and an Open Source Software Guideline[9] among other projects. All these projects are hosted as repositories on

[2] BC Developer's Exchange (https://github.com/BCDevExchange)
[3] DevEx (https://github.com/canada-ca/devex)
[4] Web Experience Toolkit (https://github.com/wet-boew/wet-boew)
[5] Open First Whitepaper (https://github.com/canada-ca/Open_First_Whitepaper)
[6] Open Government Licence – Canada (https://github.com/canada-ca/open-source-logiciel-libre/blob/master/LICENSE.md)
[7] Open Source Advisory Council (https://github.com/canada-ca/OS-Advisory_Conseil-SO)
[8] Open Source Digital Playbook (https://github.com/canada-ca/digital-playbook-guide-numerique)
[9] Open Source Software guideline (https://github.com/canada-ca/open-source-logiciel-libre)

GitHub, but they are mainly driven by the TBCS to encouraging the use of open source software or open collaboration.

Another strength the participants found about GitHub was its project management system, even though it is "a little bit of a lightweight project management system, it is still pretty good" (Participant B). It includes features such as: tracking changes from users, flagging issues, revision history, branch management, and privacy control. Git is designed to log every change made from any user, which is utilized by GitHub when it displays this information on their platform. Showing this information to all members of the organization significantly improves collaboration because transparency is a way to hold developers accountable for their contributions. As Participant B states, "a great command is 'git blame' to see line by line who did what and when. […] Like 'What? Who wrote this? Why?', you type that in and you know who to ask instead of sending out an email to your group and now 20 people have to read it, 19 of them say it wasn't me and […] it just wastes a bunch of time". If there are any issues with the code that any user has contributed, then specific changes can be rolled back to mitigate issues reactively or to block new changes proactively. Once issues are identified, they could be tracked by flagging issues on specific lines of code or creating discussion posts in an online forum in the repository.

Git also allows branch management where users can fork code into new branches and work on an alternate version of the project repository until it is ready to be merged with the master branch of the repository. GitHub allows all branches of a repository to be displayed on their website which can be used by project managers to test out the new changes with stakeholders before deciding to merge the experimental code to the main version. For example, Participant B states "you can see what a group is doing because they would deliver on a different branch", allowing multiple users to work simultaneously on the same project for different reasons. GitHub also offers privacy controls where a project can be made publicly visible or private to the organization members and project collaborators. The rights to view or edit files, or entire branches can be controlled on for each individual user.

Although GitHub offers a significant amount of project management features, not all participants felt their organizations were using them all at once. Participant G stated that they “started out with blank, many of us having never used it. First it was like dump your code here, then version control, […] you start adding issue tracking and bug tracking. […] We’ve been expanding on features as we get comfortable”. As a project would mature and requirements may change, users would try out the features they felt would best help them in their projects. Overall, the combination of Git and GitHub to allow open collaboration, version control, and project management that is specialized for software development was viewed as a big improvement over other general-purpose collaboration tools or software bug tracking solutions. As Participant B states, “you can’t use Google Docs or something to write code together or it would be a disaster” because it isn’t designed to handle computer software code or data as well as GitHub. Google Docs is a general-purpose collaborative software for office files, whereas GitHub is specialized for computer code and software file types.

Lastly, participants found that GitHub had a very low barrier to entry. Git, the underlying version control system is a free and open source software that can be downloaded and installed on any user’s computer for them to make a version-controlled project that can be hosted on GitHub for collaboration. As Participant B states, “the thing about Git is that you don’t need permission from your organization to use it”, all you need is an Internet access to download Git, and the administrative privilege to install it. As for GitHub, it has a website that can not only can display all the files of a project repository on its website, but it can also allow users to make edits to most text-based files types (HTML, JavaScript, Python, etc.) directly through their website. Since GitHub is accessible through their website, government workers don’t need to go over too many hurdles if they intend to participate in a project hosted on GitHub. As Participant E states, “it is available to pretty much every Government of Canada department, and that is a big factor because not every department has access to the same tool. […] People can get accounts and log in, […] it is publicly accessible, […] it is free to use, and you don’t have to worry about any procurement issues with getting our own web server”. Currently, not all government departments allow for the same level of access to websites, download files from the internet, or to install software that is not provided from the government's Information Technology (IT) departments. If government policies were to change, then even the access to GitHub's website can be blocked if the use of the platform is deemed unsuitable for the organization. Recently however, the Government of Canada had released the 'Directive on Enabling Access to Web Services: Policy Implementation Notice'[10], which Participant D explains is "basically a message to all IT departments saying that you need to start unblocking stuff". Additionally, the Government of Canada had released a 'Policy on Management of Information Technology'[11] that under section 6.4.9 requires the Chief Information Officer of the Government of Canada and the Secretary of the Treasury Board to "establish guidance to support innovative practices and technologies, including open source and open standard applications, and agile development". Overall, these changes should encourage government workers and departments within Canada to adopt modern, open source, and open collaborative solutions for their projects.

2.3.2 Weaknesses

The most significant weakness to the use of GitHub that participants expressed was the challenge to learn version control systems such as Git. As Participant F states, "a lot of the staff that come in, would come in with the business skills, but they don't have any of the technical skills, for example how to use version control. Pretty much every new staff that comes in, I have to show them what version control actually and that is a huge challenge". Git is a software language that requires specific keywords to be used in specific formats in order to execute its commands and operate the version control system (Blischak, Davenport, & Wilson, 2016). Participants expressed that although the software developers who regularly use Git know how to operate it, non-developers who occasionally contribute changes find it a challenge to remember those commands. Even going beyond the Git software, the

[10] Directive on Enabling Access to Web Services: Policy Implementation Notice (https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32588)

[11] Policy on Management of Information Technology (https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=12755)

collaborative software development practices of branching, pushing, pulling, or merging code is a challenging concept for non-developers to understand. As Participant F states, "one [drawback] is the learning curve, especially for new staff which creates a bit of an issue because it is already quite a learning curve in our field (data science), so just to add one more element to that learning curve is always a difficult decision to make". Adding the need to understand version control systems on top of a worker's other responsibilities can make their jobs more challenging or increase the cost to train staff in order to use open collaboration.

For new departments or teams using version control systems for the first time, the technical understanding of version control systems became a challenge when there was no support to learn it from their IT department or other departments. As Participant E states, "there is nobody to help you if you don't know what to do. We were not getting IT help to do that, nor much support for it. […] We were able to [manage] because we have people here who are more experienced, but we wouldn't expect one of our partners to do that". Some participants expressed that in projects which were in partnership with external stakeholders, they would need to re-think how to appropriately share their work with their partners. When sharing content through GitHub, besides just looking at the files on the website, external partners would need to learn how to pull files from GitHub, and merge changes into the master branch. For private projects, they would need to setup the appropriate security access for their GitHub account in order to see it, making things more inconvenient than just emailing a file to someone.

When it comes to projects that are designed to become open source projects, some participants expressed that they face challenges in sustaining their projects which mainly stem from how the data is setup, and who oversees the project. Participant B stated that "you can't just put all of your code online. You need to make it easy to start a project. You can't just copy a bunch of files […] if it is a mess then the public can't work on it". In order to be useful, open source software projects need to have adequate orientation for new comers which can be in the form of providing documentation on 'how to contribute', current issues, list of bugs, and finding a task to start with (Steinmacher, Conte, Gerosa, & Redmiles, 2015). As for managing an open source project, software developers may not be the best individuals to lead an open source community. Although software developers may be excellent at writing code for the project repository, there is a strong need for documenting the project, tracking issues, assigning responsibilities, and other skills that are better suited to give to a person with business skills who can act as a project manager. To allow for full open collaboration, there needs to be staff that oversee the contributions of all the developers and figure out how to sustain the project in times of conflict over which features should be prioritized for development. In a traditional open collaboration setting, it is difficult to figure out who is responsible for managing what part of the project because it is nobody's dedicated job to fix things (Steinmacher, Conte, Gerosa, & Redmiles, 2015). The organization needs to build a culture of collaboration and distributed responsibility. In addition, good documentation of the project can also lead to it better collaboration within stakeholders of the project, as well as discoverability of the project for others.

Another major consideration expressed by the participants was how to compensate developers for their time when contributing to an Open Source government project. Regardless of whether the developers are staff members of the government organization, external contract workers, or members of the general public working in their spare time, compensation for their work needs to be considered even for open collaboration projects. Another aspect that needs to be considered is ownership of the IP that is generated by all the contributors of a repository. It needs to be stated whether the IP contributed by the user is owned by the user or the organization that controls the project. One workaround that some participants say their organizations used was that they only considered contributions from users who were committing changes from a GitHub account that was registered to an approved government email. Users were discouraged from making contributions from accounts made via personal email. Although this method can be used to safeguard the development and ownership of the IP, it restricts open collaboration to only internal government employees, and not a broader collaboration with the public.

Some participants also expressed that GitHub was limited in accessibility because the platform was only available in the English language. Participant D stated that "in the Government of Canada, we have strict policies we need to follow. Especially for official languages, everything needs to be bilingual". Canada is a multilingual country that has English and French as its official languages, alongside many other languages spoken among its population. The Government of Canada is required to provide information in both of its official languages, however that obligation does not necessarily extend to a private organization like GitHub. Government projects that are displayed on GitHub are only displayed in English. The lack of multilingual support makes it a challenge for showcasing their work to non-anglophones.

2.3.3 Opportunities

The most significant opportunity that the participants saw from the use of GitHub is that it can allow governments to code in the open rather than work within their silos. Participant B states that "there is a pretty strong initiative in the Government of Canada to break down silos" because there are policies such as the Open Source Digital Playbook and the Open Source Software Guideline made by the TBSC that encourage Canadian government organizations to share their content and adopt open source software practices. Participants have experienced that their own government organization has been able to learn from other governments who have made successful projects. If a user sees a publicly available project on GitHub that they want to try for themselves, they wouldn't need to ask permission from the host organization, they can simply fork or clone the repository and try it out for themselves. Having the project repository be publicly available makes the code very discoverable. Participant G states that "I've mostly used it to steal code (within their team). We are all developing R coding skills at the same time, so if somebody has used a framework, […] then I'd want to see how they did it".

GitHub can also be used as a long-term archival tool. Since GitHub stores a record of all the contributions and issues that a project faced, other organizations can learn from the previous mistakes to not repeat them for future projects. Due to the Git version control system making a record of all changes, the history of the project can be preserved if the logs of those changes are preserved. Participant B states that "Git has this longevity to it because I held onto the history of all the files as different migrations happened. I'm not throwing away all of our institutional memory, I am actually holding onto it". It can be used as an effective archival tool because projects which are completed can stay publicly available on GitHub for as long as the controlling organization wishes. Additionally, the open collaboration nature of a project can allow developers who have moved on from project to come back for occasional changes. Participant D states that "lots of different people have contributed to the code base, so every once in a while someone who has worked on the project 3 years ago and has moved onto another job will pop over and somehow hop onto the repository and make a comment on something".

Although participants felt the advantages of increased collaboration within government departments, they felt that not all projects would justify the need for complete open collaboration. Some participants have expressed that they have moved away from GitHub because they need a stricter level of privacy to protect project data or stakeholders. Participant G states that "you have to be careful about privacy. There are some requirements that […] some types of data have to be kept within certain jurisdictions. Some types of data absolutely cannot be on GitHub, even if it is a server within Canada, even if it is a private repo". This concern mainly stems from the fact that GitHub is a private company and it is not based in Canada. Since GitHub is owned by Microsoft, which is an American company, there could be circumstances that private data about Canadian users or organizations stored on GitHub servers might have to be given over to another legal jurisdiction, causing loss of privacy and national security.

Due to data privacy concerns, some organization have switched to using GitLab which is an open source alternative of the services that GitHub provides. The main difference being that GitLab needs to be installed on a server that the organization controls themselves, unlike using GitHub's servers for free. Although it would require more staff with the technical expertise to setup GitLab and maintain a server, it could be beneficial to the Government of Canada because they can host their projects privately and work with sensitive data while being in full control of it. The Government of Canada has been experimenting with GovCloud[12], which is a project source code management service available to some government organizations to host their own projects on a protected, unified service. The strategic advantage of using a service like GitLab or GovCloud with a server within Canada is that all Canadian code, IP, or other assets would be contained within Canada. Having sensitive data about Canadians going to companies with servers in foreign countries could be been considered a data security risk if privacy cannot be safeguarded. However, the same concerns apply for the technical skill of the government staff that oversee the Internet security of GovCloud because they would need to ensure there aren't any hacks or data leaks to their servers. As for GitLab, there are reliability issues with using open source software because there would need to be assurance that security vulnerabilities within them would be fixed as immediately as a GitHub could as a paid software. Since GitLab is an open source software, in the case of security vulnerabilities, government workers could try to fix the issues themselves or pay the developers of GitLab to fix it as an enterprise service.

Overall, GitHub can change the way government organizations collaborate because version control systems have become an industry standard for software development that governments could learn to also utilize. Participant F states that "most public entities are a little behind the times when it comes to this kind of work and I think one of the reasons why adoption here is starting to pick up. I mean the acceleration of its pace in the last few years has been fantastic because someone who is formally trained in and practices software development, and the people who simply use the same tools as tools for their day to day

[12] GovCloud (https://govcloud.ca/doc/charts/gitlab/; https://github.com/govcloud)

work, that divide is shrinking”. As people get more interconnected and digitally communicate over the Internet, there is an increasing demand to work collaboratively while being distributed in geographic space. Participant G states that “we didn’t want to wait for somebody to understand GitHub or for somebody to roll out a plan like that, we just knew we needed it”. Even for departments who may not be getting the organizational support to use the platform, they can move ahead and use it anyways because of its low barrier to entry of being available through the Internet, and to utilize code from other organizations on the platform.

2.3.4 Threats

Some participants felt that the most significant threat to the use of GitHub for government work was that GitHub is a proprietary software that is owned by Microsoft, and all the data from Canadian government organizations is not guaranteed to stay within Canada. Additionally, Microsoft could change their monetization strategy by changing the payment structure they currently have, possibly making it more expense to use the platform than what some government departments can afford. GitHub was acquired by Microsoft on June 4, 2019 (Wanstrath, 2018), and soon after Microsoft announced on January 7, 2019 (Friedman, 2019) that GitHub will allow all repositories to be privately available for free, changing the monetization focus from private hosting of code to enterprise services instead. Some participants expressed that this may be a temporary strategy to get its current users to stay with the platform, but they are unsure if Microsoft may change its pricing in the future. Participant G states that “when Microsoft bought GitHub, it was still very easy for me to transfer my stuff from GitHub to GitLab. My repo went out, all my issue tracking, my bugs, my personal projects, all the records and history, it seamlessly transferred over to GitLab with their APIs. It was all open and great, but that will not remain the same. You give it two years and GitHub will have some stupid magical upgrade and it will no longer allow GitLab to absorb projects like that”.

What extends from GitHub being a private company outside of Canada is that there is always a threat that sensitive Canadian data could leave Canadian jurisdiction. Participant F states that “we do work with proprietary datasets and in some circumstances, we use proprietary or confidential data which cannot be distributed outside of the organization, nor can it be accessible outside the organization. […] We have to balance the fact that we have those needs for confidentiality with the risk of anything that we do getting out”. Although it could be tempting to use all of GitHub’s features to store repository files or datasets, governments need to make serious considerations about who’s data they are storing and where in order to prevent a breach of privacy or leak sensitive data.

Another threat that some participants experienced was a lack of policy coverage and understanding from Canadian federal regulators about what software is suitable for governments use. The Government of Canada has only loosely shown support for open source software. For example, the Directive on Management of Information Technology[13] which is a guideline about federal government technology mentions in Section C.2.3.8.1 to use Open Standards and Solutions by default “where possible” but it does not define what that means. Not only Git, but the use of any open source software can change if the policy does not appropriately define if the organization should trust it enough to use it. Some participants mention that due to Statistics Canada’s Section 6 about Threat and Risk Assessment[14], there are times when Open Source software gets rejected because sometimes developers are not able to respond quickly to security issues in their code.

2.4 Discussion

The findings from this chapter add to the considerations that government organizations need to make about online platforms that allow them to become open

[13] Directive on Management of Information Technology (https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=15249)
[14] Statistics Canada – Threat and Risk Assessment (https://www.statcan.gc.ca/eng/about/pia/generic/section6)

governments, or to engage in open collaboration. Mergel (2015) studied the use of GitHub in government organizations within USA had found that there was a growing usage of the platform within America but questioned how the reuse of code could lead to open government initiatives in other countries. Longo & Kelly (2016) took an early look of GitHub in the public administration in Canada and discovered that its usage within Canada was limited and driven by those with high levels of familiarity with the platform and its technologies. The results from this chapter show a matured look of how government organizations within Canada are adopting the platform as more government workers are hired with the expertise to use GitHub, or more departments are training themselves to use these technologies. Governments are catching up to modern industry standards for collaboration that depend on third party platforms.

Online platforms can allow government organizations to follow through on their initiatives to become an open government without having to spend resources on the development cost of building or maintaining open collaboration platforms. By using already existing platforms such as GitHub or GitLab, governments do not need to develop their own proprietary platforms and can focus on the innovations made from the collaboration rather than facilitating the collaboration process itself. Third party platforms can help government organizations fast track their progression in the stages of the Open Government Maturity Model (Lee & Kwak, 2012). Having transparency of data and public participation could be as simple as just making a project repository be public instead of private. Open collaboration wouldn't just be limited to the citizens of the government, but all users of a platform who wish to contribute to the government's work. Users could even fork a project and take it in their own direction if they can implement it for a different use case, or they can develop a better solution than the current users working on it.

Beyond just the platforms such as GitHub, open collaboration as an approach to create data or innovations of public value is becoming a viable option as more people become digitally connected via the Internet and educated on how to use these platforms. Open collaboration is a low-cost approach to gather a large volume of users to crowdsource innovations that can help everyone in the end if shared back as an open source software or open data. Open collaboration allows all users with the opportunity to collaborate on innovations that affect them, or at least participate in the process by voicing their thoughts. Open collaboration approach does not need to be limited to computer software development because it can be used to create innovations that are meant to be used by everyone.

Considering the weaknesses and threats to the use of GitHub for government work, a common theme is in the lack of understand of how these technologies work. Although Git and GitHub are well established technologies in the realm of software development, government organizations often lag the private sector in adopting these technologies. Even for the open government or open source software policies that governments follow, some policies need to better define what software is appropriate for government use. For government organizations to adopt open collaboration platforms like GitHub, the challenge is to gather a critical mass of government workers who are educated in modern tools and techniques of open collaboration. Knowing how these technologies perform and what policies they adhere to can allow governments to make better decisions on what software is appropriate to use for working with either open source or private projects.

2.5 Limitations

Although the total number of participants for this study was eight people and may be considered small, the participants represented their respective departments and shared the perspectives and insights on behalf of their department unless they explicitly stated otherwise. Additionally, all the participants were experts in the usage of GitHub or at least advocates for its usage for government work. The segment of government workers who were interviewed fall within a small niche of government workers who participate in open government initiatives, open collaboration, and specifically use GitHub. It is possible there are other open collaboration platforms, or open government initiatives, or government staff who could be studied if GitHub or another online platform was the focus of the research.

Although the number of interview participants was small, it is possible that given enough time and outreach to government organizations it would be possible to get more participants, especially from governments in the Western regions of Canada. The current interview participants were mostly from Eastern Canada, specifically from the City of Ottawa and Toronto. There are other Canadian government organizations such as the Province of British Columbia, Province of Alberta, and the City of Montreal who are also frequent users of GitHub, but unfortunately there were no interview participants from them. For future attempts at studying this topic, a short survey could also be used in addition to the phone interviews. A survey could reach a broad group of government workers who have used GitHub or other open collaboration platforms, and at the end of the survey ask them if they would also like to participate in a phone interview for an in-depth discussion.

The scope of this research was limited to studying the usage of GitHub, however some of the participants said that their departments were switching over to using GitLab instead of GitHub. If the scope of this research was expanded to any open collaboration platform used in government organizations, it could expand the number of interview participants and give a broader view of the use of open collaboration. However, having more participants may might not significantly change the overall themes identified in the results of this research.

Chapter 3

3 Mapping Open Collaboration of Open Government on GitHub

3.1 Introduction

Social coding has led to a notable shift in the way software is developed by utilizing open collaboration, crowdsourcing, and open source software instead of relying on proprietary or closed source products. It requires a distributed set of collaborators with varying levels of knowledge about a topic, commitment to participate, and social capital within their community (Baldwin & Von Hippel, 2011; Dabbish, Stuart, Tsay, & Herbsleb, 2012). Websites like GitHub can allow users from any background to collaborate on software or participate in online software communities. With over 23 million user accounts and 3 million organization accounts (GitHub Organizations, 2019; GitHub Users, 2019), GitHub is a significant driver of open collaboration for open source software that is used by the public, private organizations, and government organizations (Longo & Kelley, 2016). Social coding has become a field of research that aims to analyze a social network by adopting traditional network analysis techniques (Surian, Lo, & Lim, 2010; Lima, Rossi, & Musolesi, 2014). Although the data from GitHub has been used to analyze user collaboration patterns between users (Yu, Yin, Wang, & Wang, 2014), little research has been done to understand the collaboration patterns between government organizations on the platform (Mergel, 2015). The following sections explore previous literature on how open collaboration was analyzed as a social network, and how the analysis techniques can be reframed for analyzing governments organizations using GitHub.

3.1.1 Social Network Analysis

Otte & Rousseau (2002) proposed 'social network analysis' as a new field of research for information sciences by utilizing traditional network analysis and graph theory to analyze networks of social entities. Social network analysis is an approach to quantify the social structures between individual actors by primarily focusing on the relations between the actors, and secondly on the properties of the individual actors (Otte & Rousseau, 2002; Knoke & Kuklinski, 1982; Wellman & Berkowitz, 1988; Scott, 1988). For example, a social network can be constructed from a set of academic authors who have written articles themselves or co-authored articles with others. The number of articles written and who writes them can be used to quantity an author's social standings and influence they have on their social structure. Social network analysis was also used by Stein, Kremer & Schleider (2015) in order to show how the data on OpenSteetMaps was generated in an open collaboration between users on the platform. They made a co-authorship graph of users editing content in order to quantify the influence of users on the platform, and the clustering of which users collaborate more closely.

The social network analysis techniques mentioned in this sub-section are sourced from the Otte & Rousseau (2002) article, however the specific techniques originate from previous literature on graph theory and network analysis. For example, the literature includes works from 'Social Network Analysis' by John Scott (1998), 'Social Network Analysis and Education: Theory, Methods & Applications' by Brian V. Carolan (2013), and 'A Set of Measures of Centrality Based on Betweenness' by Linton C. Freeman (1977).

A social network can be represented as a graph of nodes and edges. Nodes represent any entity, while the edges represent the connection of any two entities based on some form of interactions between them. Additional data can be added to the nodes and edges to give them weight or supplementary attributes. A network of nodes and edges is typically undirected, but directions can be added to the edges to represent the influence of one node on another. Any two nodes have a path if they are connected to each other by adjacent edges, or through a series of nodes and edges in between them. A component of a graph is a subset of the entire graph where all nodes within it are connected to each other via paths between the nodes. The graph can have stray or disconnected nodes which have no connections to any other nodes. If the entire graph can be represented as a component, as in all nodes are connected to all other nodes via edges or paths, then the graph is a connected graph (Otte & Rousseau, 2002; Scott, 1988).

Social networks can be further analyzed using graph theory by calculating properties such as density, degree centrality, closeness centrality, betweenness centrality, average shortest path, and diameter. Density of a graph represents the measure of how connected the nodes in the graph are. If every node in the graph is connected to all other nodes via edges, then that graph is considered to have the highest density and be a complete graph. All centrality measures generally represent how well connected, or how important any node is within the entire graph. Degree centrality of a node is the measure of how many adjacent edges that a given node has. Closeness centrality of a node is the measure of the sum of the distances a node has to all the other nodes based on the shortest distances between them. Betweenness centrality of a node is the measure of how many shortest paths pass through a given node. The average shortest path of a graph is the average length of all the shortest paths between any two nodes. Lastly, the diameter of a graph is the largest of the shortest paths between any two nodes in the graph. The diameter is calculated by finding the shortest path between all nodes and selecting the longest of the path distances (Otte & Rousseau, 2002; Carolan, 2013; Freeman, 1977).

3.1.2 Social Coding Between Users and Organizations on GitHub

GitHub uses Git, a version control system, to log all versions, commits, and user interactions that happen on its platform. The logs of metadata generated by Git, in addition to the profile information of users and organizations, can be analyzed to better understand GitHub as a social coding platform (Storey, Singer, Cleary, Figueira Filho, & Zagalsky, 2014).

Combining social network analysis with the social coding data of GitHub, prior literature has aimed to analyze the platform for patterns in the collaborations. Lima, Rossi,

& Musolesi (2014) had built a 'Follower' graph which was a directed graph based on users following other users on GitHub. They discovered that followers who were active on the platform typically did not have many followers. Thung, Bissyande, Lo, & Jiang (2013) had built two graphs which were a 'project-project' network and a 'developer-developer' network. The project-project network was a graph of projects (repositories) as nodes that were connected by edges if they shared at least one developer in common. The developer-developer network was a graph of developers (users) that were connected by edges if they had worked on the same project. They discovered that the project-project network had an average shortest path length of 3.7, whereas the developer-developer network had its average shortest path length to be 2.47. When comparing the lengths to other human communication networks such as Microsoft Messenger with its average shortest path length of 6.6 (Leskovec & Horvitz, 2008), or Facebook with 4.7 (Ugander, Karrer, Backstorm, & Marlow, 2011), it showed that both the project-project and developer-developer networks from GitHub were more interconnected than human communication networks. Even when comparing to another software collaboration platform called Sourceforge with its average shortest path length of 6.55 (Surian, Lo, & Lim, 2010), the two GitHub networks were found to be more interconnected.

Mergel (2015) had focused on using social network analysis techniques to understand how code is shared within government organizations on GitHub. They created two networks based on the data from GitHub which were (1) a fork network, and (2) a pull request network. The fork network was a directed graph based on which government organizations were forking which other organization's repositories. Similarly, the pull request network was a directed graph based on which government organizations were sending pull requests to which other organization. They discovered that although only a few organizations were doing most of the forking, overall forking was much more common than making pull requests because it was objectively easier to fork repositories than make changes and submit pull requests. However, a few of the government organizations which were USA's chief digital services and open data divisions such as @18F, @GSA, @Presidential-Innovation-Fellows, and the @Project-Open-Data were forking just as many repositories as they were receiving pull requests, indicating that they were more accepting of open collaboration than other government organizations on the platform. Forking repositories and submitting pull requests was used as a mechanism to share code between organizations along with the control that these organizations can have on what changes they choose to accept. Considering the social coding and network analysis in the prior literature for providing a new approach to analyzing social network for information sciences, it is possible to apply these techniques to make a network of government organizations that are taking part in open collaboration on GitHub. Doing so can help uncover the social structure between government organizations and identify if there is just as much of a collaborative relation between organizations as there is between users on the platform.

3.1.3 Research Goal

The major research question of this chapter was to understand the extent and nature of government collaboration on GitHub. The specific areas of focus for the research question are outlined below:

1. Are government organizations actively using GitHub? And how many?
2. How complete is the information available about government organizations on GitHub?
3. Is there a collaborative relationship between government organizations that use GitHub?

The first research question was to analyze what proportions of the organizations were actively using GitHub based on when they last used their account, and last updated a repository. The second research question was to analyze the completeness of the information available about these government organizations. These first two questions aim to examine which government organizations are using the platform and if there gaps in the metadata about these government organizations.

The third research question was to determine if there was a collaborative relationship between government organizations on GitHub by creating a network of government organizations and applying some of the social network analysis methods. In this case, a 'collaborative relationship' exists between two organizations if any of their repositories contain a user who has contributed code to both of their repositories. This method was used instead of directly focusing on government workers because not all government workers who have contributed to a government repository were publicly listed as a part of their own government organization, making it difficult to determine if a given user is a government worker or not. Although forks and pull requests are direct indicators of whether an organization has accepted another's work, they do not explain the reason for the collaboration, and since organizations can fork as much content as they want without making any changes to it, it could inflate the number of perceived collaborations between two organizations (Lima, Rossi, & Musolesi, 2014). Any user that is listed in a repository as a contributor is known to have worked on that project at one point in time, regardless of the intention of the contribution, however an issue with this assumption is that it does not take into account the context of why a user collaborated since they could be a current member of the government organization, a retired member, an external stakeholder, or just from the public. Thus, a collaborative relationship is only considered between two government organizations if they share a user in common, and not focus on the social or political context of why the individual user collaborated, but rather focus on the organizations that are collaborating with each other. The resulting network was also plotted on a geographic map to visualize the connectivity between the organizations worldwide.

3.2 Methods

To answer the research questions, data was collected from GitHub organizations that were listed on the Government GitHub List (GGC) list, specifically information about the organization, their repositories, and all users who have contributed to them. The GGC list was used because when searching the GitHub platform by just the keyword 'government', over 1,600 accounts[15] are recommended and only a few of the results are actual government accounts. On the other hand, the GGC list is a crowd sourced list of government accounts made by openly asking users to add new accounts names they believe are government accounts. The collected data was used to answer the first two research questions, and to create a network to answer the last research question about understanding the collaborative nature of the organizations. The following subsections discuss the methods used to collect the data from GitHub about government organizations, as well as the creation and analysis of the organization network.

3.2.1 GitHub Data Collection

The data collection workflow outlined in Figure 8 was conducted using Python libraries in a Jupyter Notebook. The first step involved using GitHub's Rest API to make a series of web request for data about specific GitHub organization accounts listed in the GGC list. As the GitHub API would answer the request with the data on the GitHub accounts, they were stored as a table using a Python library called Pandas. For each organization account, there was information about them such as their name, description, contact email, as well as a geographic location of where the organization was in the world. Since the location data on these accounts was in plain text, it had to be geocoded, meaning that it had to be translated from plain text and into latitude/longitude pairs which could be used to place the organization accurately on a world map. After using Google Maps API to geocode each account, the geospatial dataset was stored in Microsoft Excel spreadsheets.

[15] GitHub account search results for 'government' (https://github.com/search?q=government&type=Users)

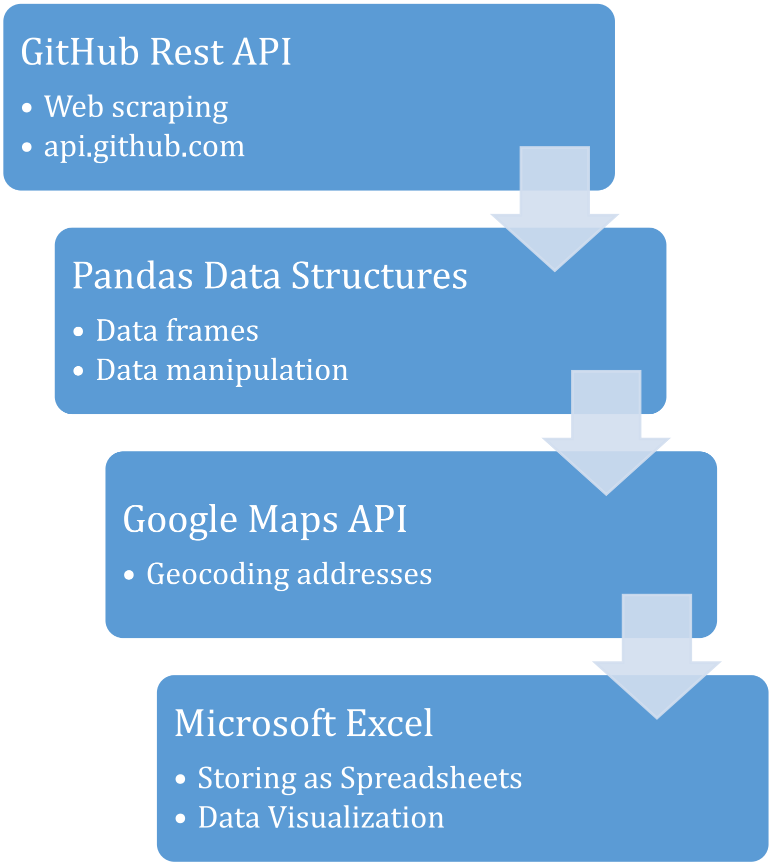


Figure 8: Data collection workflow

3.2.2 Data Structure

After completing the data collection, a dataset of was made consisting of four tables (1) countries, (2) organizations, (3) repositories, and (4) contributors. Countries table represented 59 countries from the GGC list which had government organizations using GitHub. Organizations table represented 782 government organizations from those countries, with information about their name, description, location, contact information, public members, data of creation (when the account was created), and date of update (when the account was last updated). Repository table represented the 31,703 public repositories that were owned by the government organizations, with information of which repository belonged to which organization. The contributors table represented a list of users on the GitHub platform, and information about which repository they contributed to. For the tables of countries, organization, and repositories, each row is a unique entity in the table, i.e. no duplicate of those countries, organizations, or repositories can exist. Although the data from forked repositories could be identical to their parent repository (when no changes have been made to the fork branch), they are treated as a unique repository and adds to the number of repositories to the organization which forks it. As for commits table, it contains duplicates of users because one user can commit to multiple repository owned by multiple organization. What makes the commits table unique is the combination of user identification (ID) and the repository ID to make it a unique interaction between a user and an organization based on which repositories they interacted with. In this case, there are 227,322 interactions, coming from 71,549 unique users who have contributed to government repositories.

3.2.3 Metadata Analysis

In order to determine how many of the government organizations are actively using the platform, information about when the accounts were last updated, or when their repositories were last updated were plotted on a graph. If the organizations are actively using the platform, they will have their accounts and repositories be updated as recently as possible. Every account and repository on GitHub contained metadata about when it was created, and when it was last updated. Although GitHub considers an account updated for any change made to the account or any of its repositories, for the context of this research the date of when an account was late updated will be taken as an overall representation of when any activity on the account happened to indicate that it is being used or not, and the research does not make a distinction on the quality of the work being done by the organization.

Additional metadata representing contextual information about the organization such as name, description, location, email, and more were plotted on a graph to examine what proportions of the organizations provided such information. Providing complete information about government organizations improves the credibility of their organization on the platform because such basic information is needed to have initial conditions and data transparency in order to become an open government (Lee & Kwak, 2012).

3.2.4 Network Construction

Using the information from the commits table, specifically the information on which user contributed to which repository, a graph was made to represent the network of government organizations collaborating on the platform. This research assumed that two organizations were collaborating if they shared at least one user who had contributed to both of their repositories. Meaning, that if a user ID was listed as someone who made commits to an organizations' repositories, then that organizations was considered to collaborate with any other organization that allowed that user to also contribute to their repositories as well. This method was used instead of creating a directed graph based on the forks and pull requests as done by Mergel (2015) because this research assumed that any user who was allowed to contribute to a repository was participating in open collaboration, regardless of the purpose, quality, or quantity of the contribution.

Assuming two organizations are collaborating if they share contributors has a few edge cases where the assumption may not apply if every commit were to be analyzed for the intent or purpose of the collaboration. There are cases when two organizations will look like they are collaborating, but that may not be their intention. It is possible that two organizations are not planning on collaborating, and a user happens to work for one organization and occasionally writing code for others as open collaboration. It is also possible that users could have written code for one organization, and changed jobs to write code for another, which would make it seem like the two organizations that they worked for had collaborated. It is also possible that some government organizations hire third party consultants or software developers to contribute to their repositories, and any work that those external developers had done with other government organizations will also show in the network as a collaboration between two government organizations. Lastly and most commonly it is possible that a government organization has little to no content that is originally developed by them and they have mostly forked repositories from other organizations, which will make it seem like they are collaborating with others significantly more than others.

These edge cases bring about some considerations about the open collaboration nature of GitHub. The intention of open collaboration is to allow users to freely contribute to any organization on the platform, but it is difficult to assess the purpose and quality of a contribution that a user makes on the repository of an organization in order to make a distinction on which organization a user is working for. It is possible to use additional metrics of a user such as the number of lines of code written, number of commits made, or number of pull requests accepted, but these metrics could be subjective in themselves because they do not represent the quality of the code. Hence, this research does not consider the quality of code written by a user for any government organization, but only that they are listed as a contributor to the repositories of an organization.

In the graph, the nodes represent the organizations on GitHub, while the edges represent a connection to other organizations that they have collaborated with based on which users have worked on both of their repositories. The edges also have weight to them which is the total number of users they share. Any stray nodes that have no connection to any other organization were removed from the network resulting in the final graph being a connected graph.

3.2.5 Network Analysis

The organization network can be analyzed by calculating graph properties discussed earlier which include the centrality measures, average shortest path, density, and diameter. If the graph has a high density, meaning if most of the nodes are connected via adjacent edges, it would indicate that most government organizations are open to collaboration with each other. Degree centrality of nodes with the most adjacent edges, as in organizations with the most connections, can be considered as organizations most open to collaboration because they accept input from many users on the platform who have also contributed to other organizations. Edges with the highest weight, as in pairs of organizations with the most shared users can be considered to have a strong collaborative relationship with each other because they have the greatest number of users who have contributed to both of their repositories.

Closeness centrality and betweenness centrality indicate which of the organizations are at the center, or influential in the network because they are either close to all other organizations, or they are between organizations that may collaborate with each other. The average shortest path and the diameter indicate the overall size of the network of organizations collaborating with each other. Having a small diameter and average shortest lengths values signify that the network is compact, and that organizations have a lot of overlapping sets of shared users. These network measures can help determine if there is a collaborative relationship between the government organizations, and the influence these organizations have on the entire network of collaborations.

3.3 Results

The results of the analysis are grouped into the following subsections that answer the three research questions respectively.

3.3.1 Active Government Accounts

After collecting data on all the 782 government GitHub accounts listed on the GGC list, they were plotted based on their data of creation. Since the creation of GitHub in 2009, its adoption has increased year over year until 2014 where that increase plateaued (Figure 9).

Although the increase of new accounts has plateaued, this plot does not indicate whether all the accounts that have been created are active or not. Additionally, a single government body could have ownership and control of multiple GitHub accounts and information on the ownership of the accounts to their real-world government organization is often not listed and rather implied by the name of the account. For example, the account @thecityofcalgary is owned by the City of Calgary in Canada, but the accounts @canada-ca, @web-boew, and @cds-snc are owned by the federal Government of Canada.

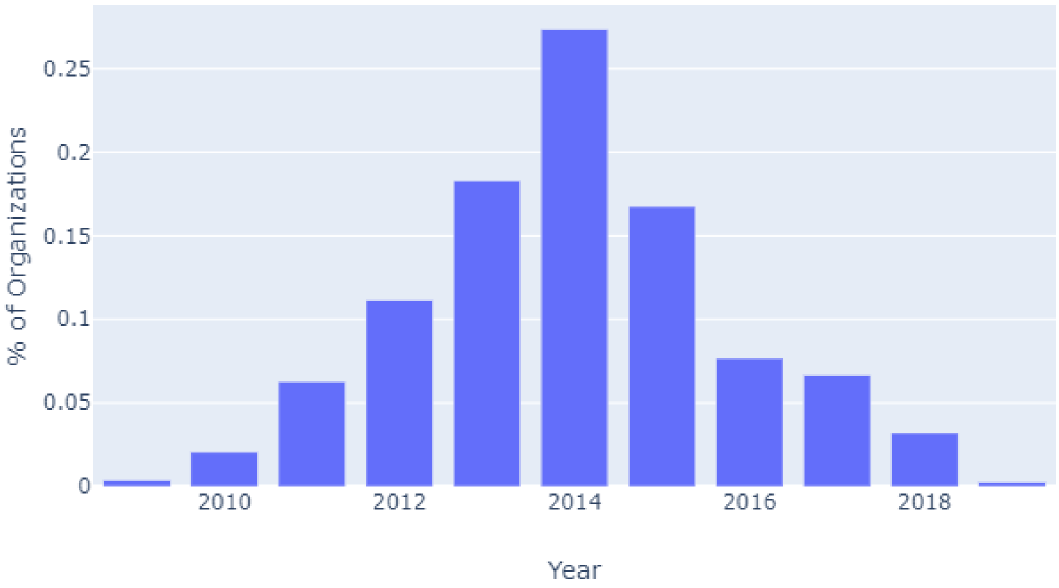


Figure 9: Percentage of GitHub accounts by their year of creation (N=782)

In order to understand how many government organizations were actively using the platform, they can be plotted based on when their account was last updated. Figure 10 illustrates the proportions of government accounts by the year of when they were last updated. It shows of the 782 organizations, 28% and 26% were updated in 2018 and 2019 respectively, and that over 50% of the accounts were updated within the past two years. However, there are some that have not been updated since 2014, which could indicate that either some organizations aren't using the platform anymore or they aren't using it as frequently.

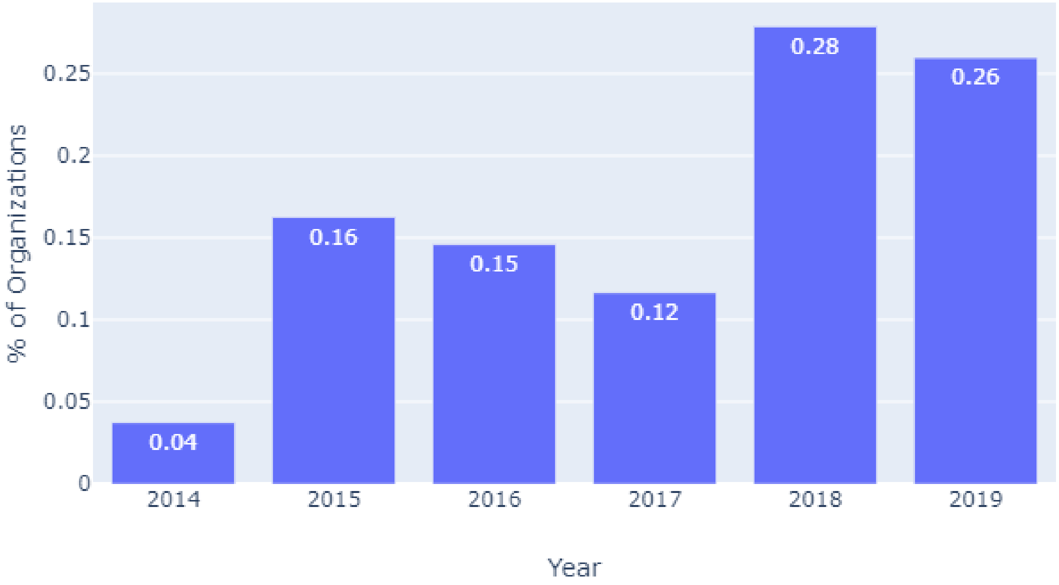


Figure 10: Percentage of government accounts by their year of update (N=782)

Not all governments function at the same pace, schedule, or manpower to be constantly active on the platform. It is possible that individual repositories from these organizations are more active than their respective organizations. Figure 11 illustrates the proportions of the repositories that are owned by government organizations and the year they were updated. Of the 31,703 repositories owned by government organization, 44% and 22% of the repositories were updated in 2019 and 2018 respectively. It also shows that up to 80% of the repositories were updated within the past few years. One thing to note is that the plot illustrates all the repositories without their association of which organization they belong to. It is beyond the scope of this research to analyze each organization for their respective frequency of updates on their own set of repositories.

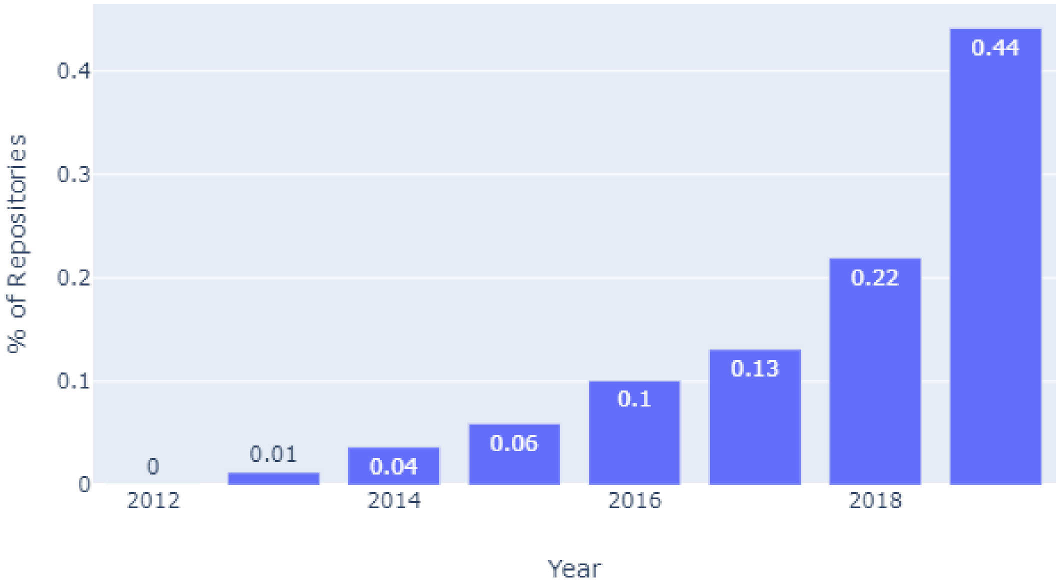


Figure 11: Percent of government repositories by their year of update (N=31,703)

From the data collected from GitHub, the countries table can be plotted on a map to visualize which countries have the greatest number of government accounts on the platform. Figure 12 illustrates that the USA is the country with the most government organization accounts on the platform (301), followed by some countries such as the United Kingdom (87), Canada (49), Australia (41), France (39), Sweden (38), and Brazil (32).

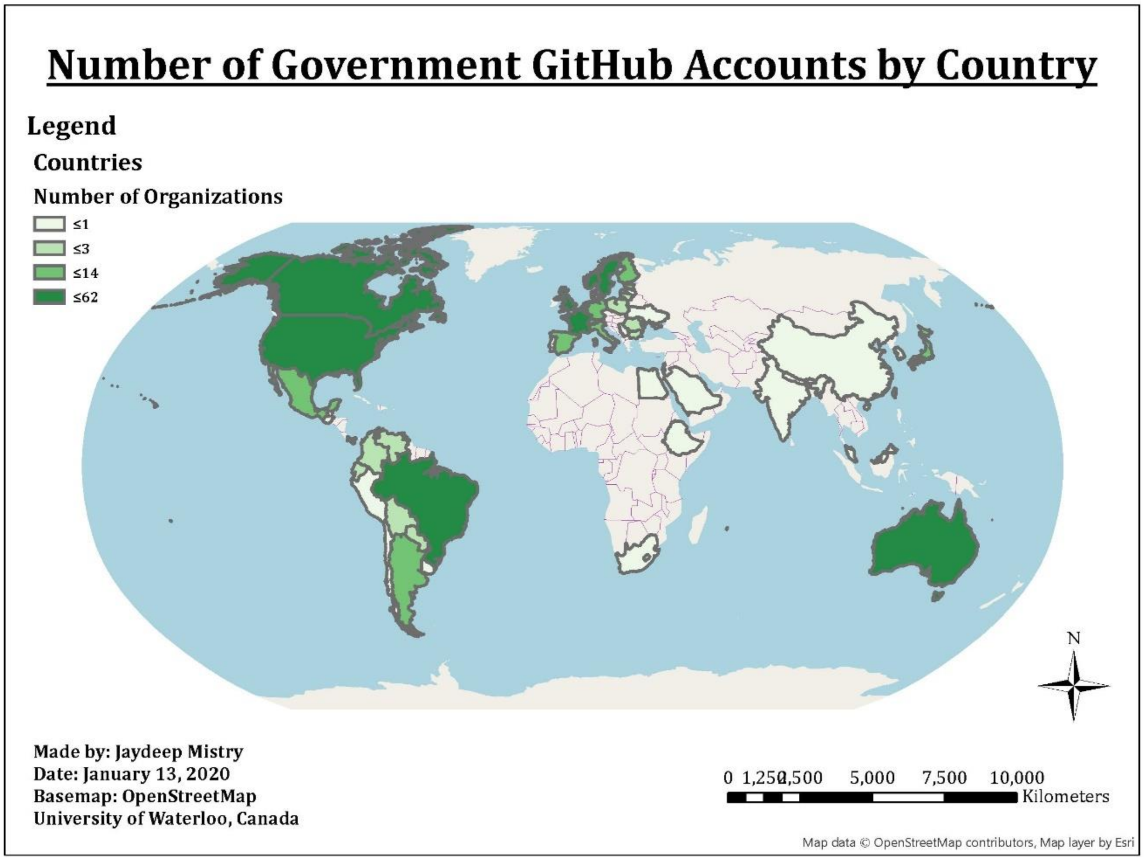


Figure 12: Number of government GitHub accounts by country (N=782)

3.3.2 Gaps in the Metadata

The second research question was to determine the completeness of the information about the organizations, which included metadata or contextual information typically required about an open government organization such as name, website, location, email and more. Figure 13 illustrates the contextual items and what percentage of organizations had provided such information about themselves. Additionally, it shows information on how many of the organizations contained repositories, members, and were verified accounts. Ideally, governments should provide all necessary information about themselves to be considered as a legitimate government organization on GitHub, however the data shows that there are many of them that do not provide such information. Less than 50% of organizations provided a description or an email to contact them. Almost 20% of them are missing a website link. Almost 30% of the organizations have no geographic location data about where they are in the world.

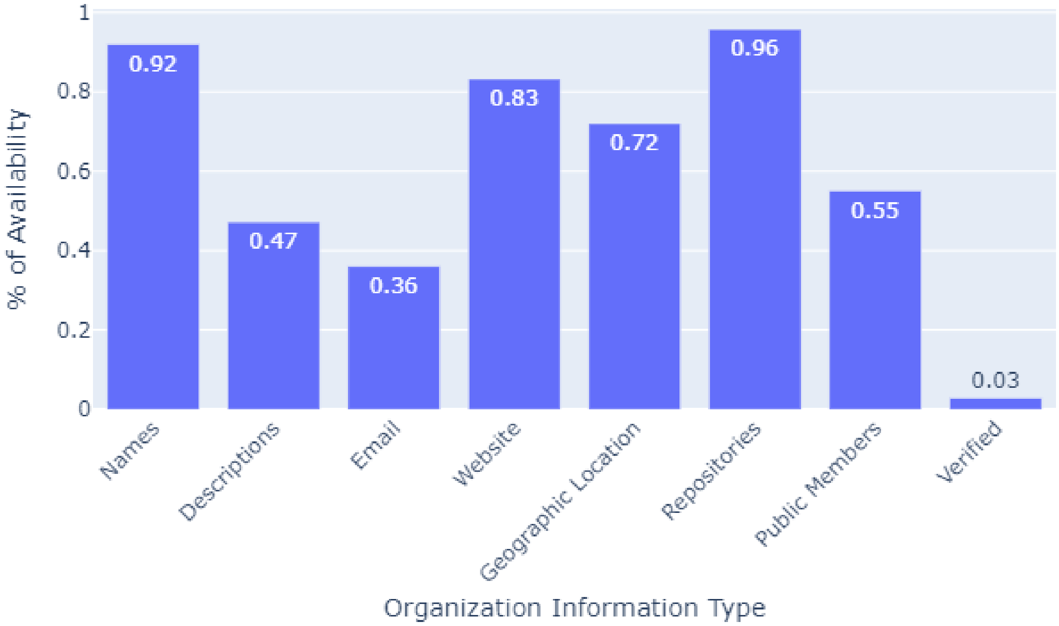


Figure 13: Percentage of accounts by their amount of contextual information provided (N=782)

Even information about who is a public member in these government organizations is often missing. Of the few that do have a public member, they are often listed as the contact email for the entire organization. GitHub has allowed the members of an organization to have to opt-in to be publicly visible as a public member of an organization. Knowing the organizations that a user is affiliated with allows other government organizations, researchers, or public members to make better judgements if they want to work with an organization. Providing contextual information about the organization, having public members, or having public repositories signifies that the organization is serious about using

GitHub to become an open government by committing to data transparency and participation from others on the platform.

Only 3% (22) of the organizations were verified organization accounts and they are listed in Table 1. These include some early adopters of GitHub as they have been using the platform since 2012. There are many unverified government organizations because the process to get an account verified versus to get on the GGC list is different. To get an account verified, the email address and website of the organization must be confirmed by GitHub to be accurate and representative of the GitHub account being verified. Whereas the GGC list is a crowdsourced list, and the only qualifications required for an organization to be added to it is to have at least one public repository and contain some reference information about the government such a website. Due to this lack of verification of whether an account is a legitimate government account or not, it is possible to abuse the crowd sourced nature of the GGC list to insert any account that is not a real government account.

Table 1: Verified government organizations on GitHub

Row	Country	Login	Name	Date Created
1	Australia	AtlasOfLivingAustralia	Atlas of Living Australia	2014-04-15 0:46
2	Australia	datagovau	data.gov.au	2013-09-21 6:15
3	Australia	govau	Digital Transformation Agency	2017-01-19 23:38
4	Canada	VilledeMontreal	Ville de Montréal	2013-11-22 18:58
5	France	ANSSI-FR	ANSSI	2012-07-04 9:13
6	France	clipos	CLIP OS	2017-11-17 14:32
7	France	clipos-archive	CLIP OS - Archive	2018-05-04 9:27
8	France	etalab	Etalab	2013-08-26 16:03
9	Norway	navikt	NAV	2015-04-08 7:08
10	Norway	Skatteetaten	Skatteetaten	2012-03-14 11:31
11	Norway	dsb-norge	Direktoratet for samfunnssikkerhet og beredskap	2017-06-23 11:29
12	Panama	miambiente	Ministerio de Ambiente de Panamá	2015-07-15 21:04
13	Sweden	SVT	Sveriges Television (SVT)	2012-03-13 13:27
14	U.K. Central	LocalGovDigital	LocalGov Digital	2014-01-09 20:12
15	U.K. Central	ministryofjustice	Ministry of Justice	2012-08-23 11:22
16	U.K. Central	ukncsc	The National Cyber Security Centre	2015-03-09 14:53

17	U.S. City	CityofSantaMonica	City of Santa Monica	2013-10-22 1:09
18	U.S. City	southbendin	City of South Bend, IN	2014-07-19 18:10
19	U.S. County	MCLD	Maricopa County Library District	2014-05-22 14:50
20	U.S. Federal	presidential-innovation-fellows	Presidential Innovation Fellows	2012-08-13 20:55
21	U.S. Federal	18F	18F	2013-12-20 23:58
22	U.S. Federal	usds	U.S. Digital Service	2014-09-12 16:22

3.3.3 Network Structure of Government Collaboration

The third research question was to examine if there was a collaborative relationship between the government organizations on the platform. Using all the data collected about government organizations, repositories, and contributors, a graph of government organizations was constructed based on shared users who made commits to both the organization's repositories. The resulting graph contains 620 nodes, short of the 782 total organizations because 162 of them were stray nodes with no adjacent edges (connections to any other organization). Although the graph contains 29,006 edges, it only has a density of 15.11%. The average shortest path length of the network was calculated to be 2.08, meanwhile the diameter of the entire network was 5.

The entire graph was geographically plotted (Figure 14) based on the geographic location provided by the organizations. For more clarity of the data, Figure 15 illustrates the top 1% percent of the organization collaboration based on the number of users they share. For the 28% of government organizations that didn't provide any geographic information, the location of their country was used to geocode them onto the world map. Appendix D contains additional geographic maps of specific organizations collaborating on GitHub which include the most active open government such as @18F (USA), @alphagov (United Kingdom), @canada-ca (Canada), @govau (Australia), and @VilledeMontreal (Canada).

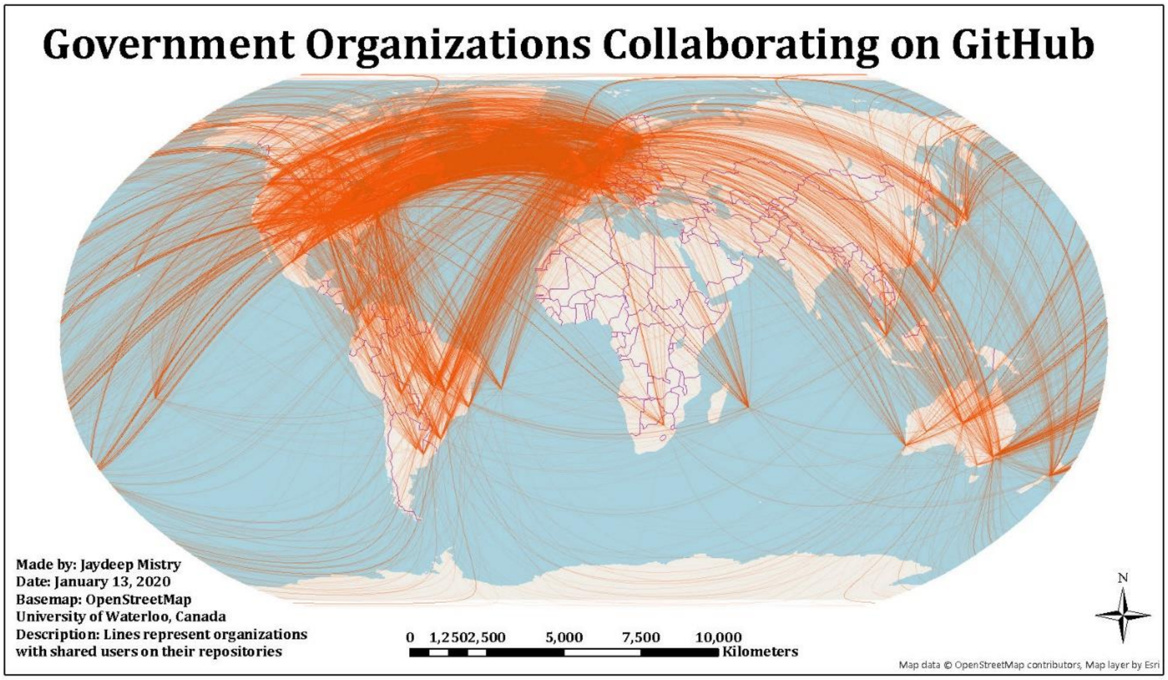

Government Organizations Collaborating on GitHub
Made by: Jaydeep Mistry
Date: January 13, 2020
Basemap: OpenStreetMap
University of Waterloo, Canada
Description: Lines represent organizations with shared users on their repositories
0 1,2502,500 5,000 7,500 10,000
Kilometers
N
Map data © OpenStreetMap contributors, Map layer by Esri

Figure 14: Global map of all government organizations collaborating on GitHub

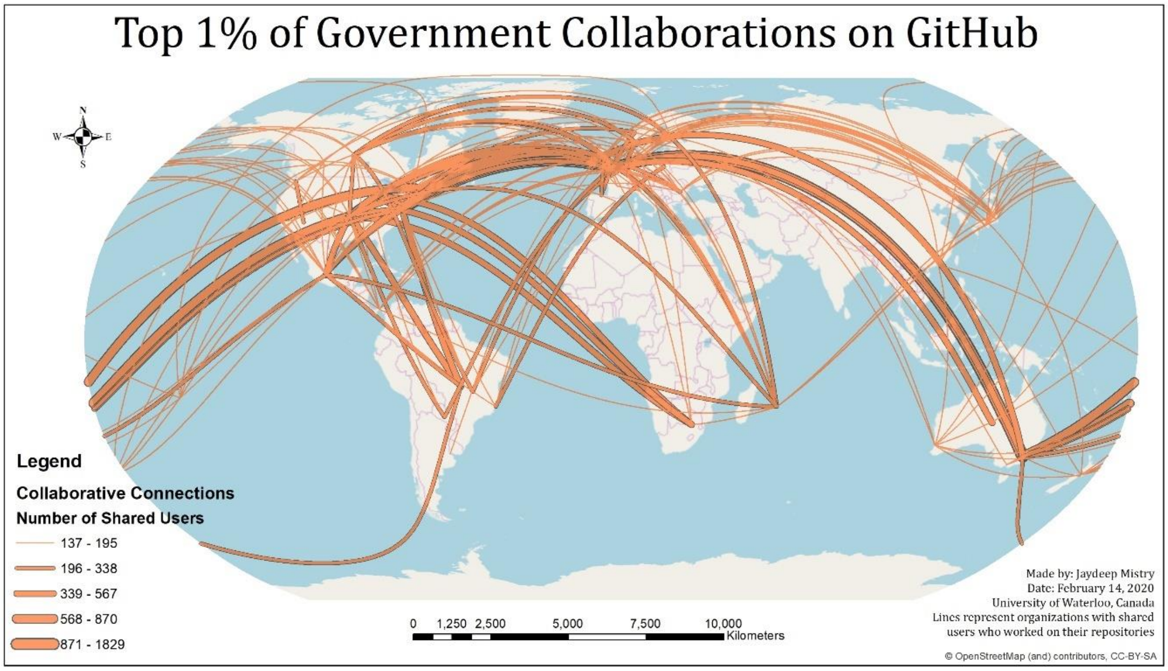

Top 1% of Government Collaborations on GitHub
N
W
E
S
Legend
Collaborative Connections
Number of Shared Users
137 - 195
196 - 338
339 - 567
568 - 870
871 - 1829
0 1,250 2,500 5,000 7,500 10,000
Kilometers
Made by: Jaydeep Mistry
Date: February 14, 2020
University of Waterloo, Canada
Lines represent organizations with shared users who worked on their repositories
© OpenStreetMap (and) contributors, CC-BY-SA

Figure 15: Top 1% of government organization collaborations on GitHub

Network centrality measures such as degree centrality, closeness centrality, and betweenness centrality can be utilized to measure the collaborative nature and influence of the government organizations on the entire network. Degree centrality is the measure of how many adjacent edges that a node has, closeness centrality is the measure of how close one node is to all other nodes, and betweenness centrality is the measure of how many shortest paths between pairs of nodes that pass through a given node (Otte & Rousseau, 2002; Carolan, 2013). Table 2 shows the top 20 accounts for each of the centrality measure. The accounts @alphagov and @18F, which are the national digital services organizations for United Kingdom and USA respectively are the top two organizations for every centrality measure. This is an indication that these two countries are also the most accepting of open collaboration through GitHub because they have the most collaborations with other organizations, as well as being very central in the entire network of these collaborations as well as sharing many users.

Table 2: Top 20 accounts per network centrality measure

Rank	Degree Centrality		Closeness Centrality		Betweenness Centrality	
1	alphagov	0.670436	alphagov	0.743097	alphagov	0.054408
2	18F	0.663974	18F	0.741317	18F	0.047912
3	GSA	0.610662	GSA	0.709862	esdc-edsc	0.042468
4	mxabierto	0.600969	mxabierto	0.706621	GSA	0.035209
5	govau	0.562197	govau	0.685493	mxabierto	0.023394
6	MUnosecc	0.542811	CityOfPhiladelphia	0.678728	govau	0.020392
7	CityOfPhiladelphia	0.541195	MUnosecc	0.671367	SSAgov	0.020062
8	esdc-edsc	0.53958	UKHomeOffice	0.669913	nationalparkservice	0.019060
9	UKHomeOffice	0.533118	esdc-edsc	0.667745	MUnosecc	0.018902
10	bcgov	0.523425	bcgov	0.667745	CityOfPhiladelphia	0.017881
11	SSAgov	0.510501	NYCPlanning	0.656416	kartverket	0.015407
12	StadGent	0.508885	StadGent	0.655720	interlegis	0.015062
13	skat	0.504039	skat	0.655026	StadGent	0.012646
14	kartverket	0.500808	SSAgov	0.654334	betagouv	0.012016
15	NYCPlanning	0.500808	dbca-wa	0.653643	simp	0.012009
16	USStateDept	0.49273	kartverket	0.652266	bcgov	0.011907
17	betagouv	0.491115	prodest	0.652266	VilledeMontreal	0.011892

18	Fedict	0.491115	betagouv	0.651579	ministryofjustice	0.011836
19	dbca-wa	0.491115	ministryofjustice	0.649528	nasa	0.011649
20	prodest	0.489499	USStateDept	0.648847	dbca-wa	0.011645

The distribution of the node degree is plotted in Figure 16, where the x-axis represents the degree of the nodes (the number of connections an organization has), and the y-axis represents the frequency of the degree (the number of organizations that have those many connections). The long tail distribution indicates that most organizations have a few connections to other organizations, whereas a few organizations have hundreds of connections to others. Although this indicates that some organizations are more open to collaboration with others, it does not take into account the proportion of contribution from the users as we only consider if a two organizations share a common user, and not how much they have contributed for one organization versus another.

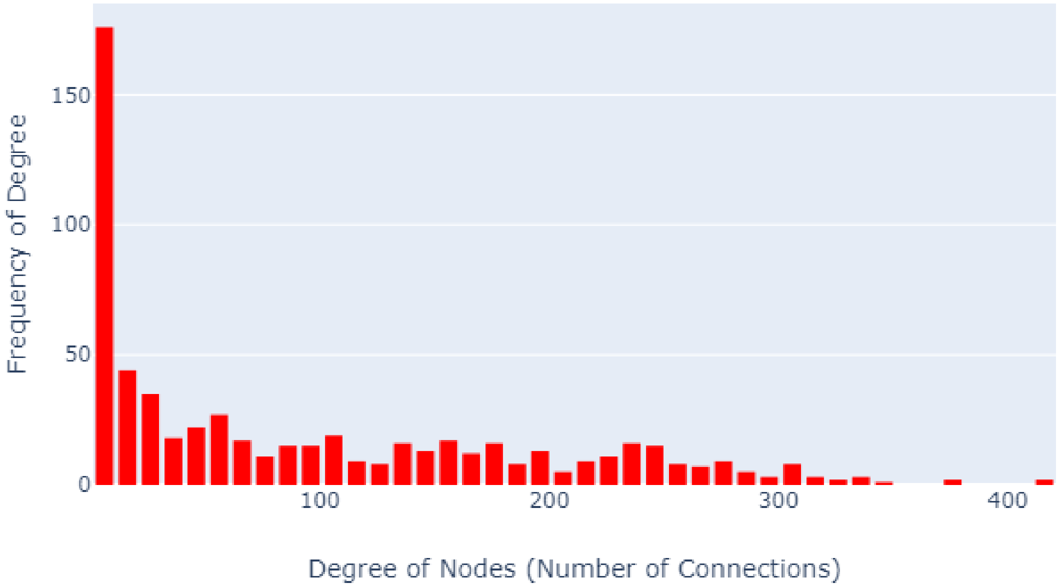


Figure 16: Distribution of node degree of government organizations (N=782)

The edges, or pairs of organizations with the highest number of shared users are illustrated in Figure 17. It shows that some organizations tend to collaborate with each other much more than others. In this case, @alphagov, @govau, and @18F, have the greatest number of shared users. Other organizations also have some strong connections such as the @GSA (General Services Administration), the @ministryofjustice from America, and even the @VilledeMontreal (City of Montreal) from Canada.

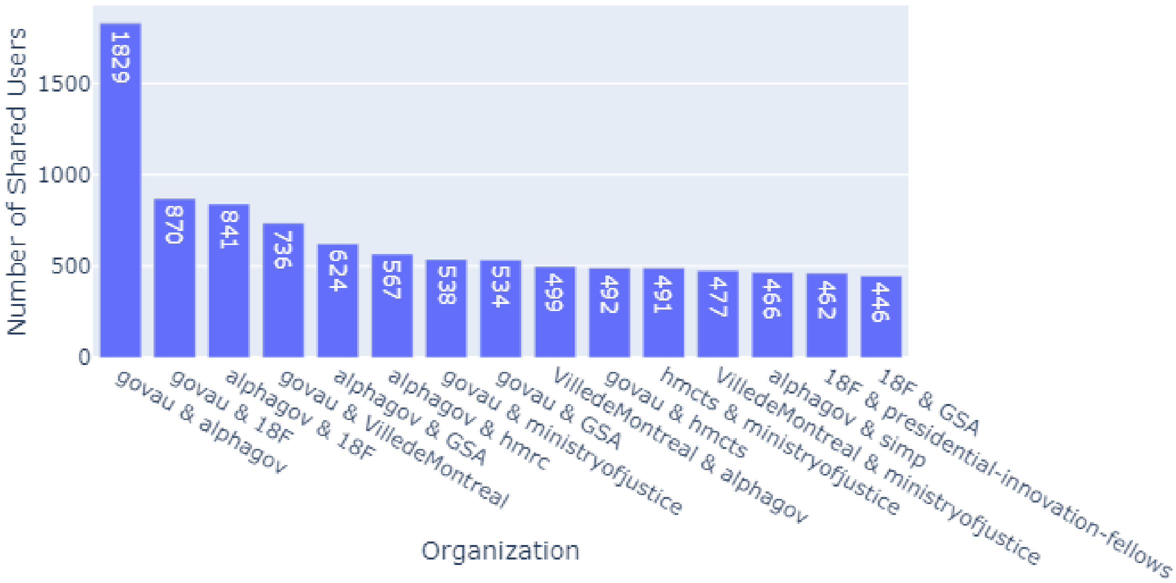


Figure 17: Top 15 edge weights from the government organization network showing which pairs of organizations are most collaborative with each other.

3.4 Discussion

Government organizations have been utilizing social coding to collaboratively create open source software and civic tech of public value. From the 782 government organizations studied in this research, at least half of them have been actively using the platform within the past two years. Although government organizations originate from all over the world, certain countries such as the USA, United Kingdom, Canada, Australia, France, Sweden and Brazil have a significantly greater number of organization accounts on GitHub than other countries. GitHub is used more in developed countries who may have policies that encourage them to become open governments via open collaborations, meanwhile its usage is very limited in most other developing countries. It could be possible that some government organizations may have been missed from being added to the GGC list, or they keep their repositories private, or they use other collaboration platforms or services. Overall most of the government organizations that do have an account on GitHub are actively using it to work on their repositories of civic tech, open data, open source software, and more.

Considering the information that the government organizations provided about themselves, there is a need for better quality control of the data they provide because it is often incomplete. Any public facing government resource that citizens want to use should provide adequate information about itself in order gain public trust (Vetro, et al., 2016). Having complete geospatial and contextual data about these government accounts can not only benefit their public perception of their data transparency, but also allow them to mature as an open government by allowing citizens to participate in their projects or collaborate with their repositories. Additionally, accurate information about the organizations can help future geospatial government research from various fields such as Geoscience, social coding, open government, open data, and more.

There is a collaborative relationship between most of the government organizations that are using GitHub because 80% (620) of the organizations have accepted commits from users who have also contributed commits to the repository of another organization. The organizations most collaborative on GitHub also belong to countries that have the greatest number of organizations on the platform, mainly from @18F (USA) and @alphagov (United Kingdom). The organizations from developed countries have very strong relationships with each other because they share many users who have worked on both of their repositories. Although the density (15.11%) of the network may not be high, the diameter (5) and average shortest path length (2.08) was small in comparison which signifies that the organizations in the network may not have as many adjacent or direct connections with each other, but they are just a few edges away from each other.

Considering the notion of “six degrees of separation” presented by Travers and Milgram (1977), they state that in a social network, any two people are separated by 6 degrees; meaning that the distance between two people is 6 based on the series of people who know each other in their network. Prior research in the field had analyzed the network sizes of much large human communication networks and found values that resemble the six degree of separation. The average shortest path length of Microsoft Messenger was 6.6 (Leskovec & Horvitz, 2008), meanwhile Facebook was 4.7 (Ugander, Karrer, Backstorm, & Marlow, 2011). In comparison, the network of government organizations collaborating on GitHub from this research is more compact than human social networks because it had a much smaller shortest path length of 2.08. Even when comparing the government network to Sourceforge, another open collaboration platform similar to GitHub, its average shortest path length was larger at a value of 6.55 (Surian, Lo, & Lim, 2010). In an open collaboration, users could collaborate with each other without having to know each other on a personal level and instead focus on the products being created (Thung, Bissyande, Lo, & Jiang, 2013).

Government organizations have been utilizing the open collaboration nature of GitHub in order to become an open government. GitHub can help governments quickly mature as an open government because all the collaborative features of GitHub encourages data transparency, citizen participation, and collaboration in the open. The open collaboration between organization can be analyzed using social network analysis techniques used to better understand their social structures.

The collaborations between users and organizations can be analyzed as a collection of nodes and edges. In this case, representing the network as a graph reveled the organizations that were most collaborative, most central to the network, and the overall nature of the relationship between the organizations on GitHub. Mapping the organizations geographically can also illustrate which countries are the most collaborative and encouraging others to work with them in the future. The knowledge of the social structure of the government organization collaborations can be used by other governments if they want to find other like-minded organizations in order to find common projects to work on or even find developers who could be employed to work on their projects.

3.5 Limitations

GitHub is a platform that is constantly in use by users who are collaborating and contributing changes onto the platform. The data collection process is time sensitive and only contains the information about these government organizations up to the point of time of when the data was collected, in this case it is June 2019. Any changes that happen after the time that the data was collected would not be reflected in the analysis, and thus the analysis is limited to the time of data collection. If the usage of GitHub for governments were to change drastically after the time the data was collected, then it would not be reflected in the analysis and the methods of the data collection and analysis would have to be redone for any future changes.

Another limitation in this study is the assumption that the GGC list is always up to date and contains the names of all current government organizations accounts on GitHub. There are a few accounts on the GGC list that have not been in use for over 5 years, but the account has not been removed from the list. There are also a few accounts that are on the GGC list that are actively used but they contain no publicly available repositories or content. Keeping all content private could be necessary when working with sensitive data, otherwise it takes away from the potential of the platform to be used to become an open government to improve transparency and public participation.

Another limitation is in the assumption that any two government organizations have a collaborative relation if they share a developer in common. It is possible that some governments workers may be contributing to other organization's projects out of their own spare time and free will, and not because their own organization has any plans or obligations to collaborate. Two government organizations could also look like they collaborate if a developer had once worked with one of them, and later changes jobs to work for the other, but the two organizations never planned to make that switch. The underlying collection of users who are collaboration on the platform and their movement from one organization to another is not considered in the research methods of this study.

Another limitation of this study is that data about current government repositories was only collected about the repositories that were publicly available. Any private content that the government organizations work on by themselves or collaborate with other organizations will not show up in the analysis of this study because only data that was publicly available was collected. Publicly collected data is specifically any information that would be available on GitHub to any user who is using the platform, without any special access or privilege to see any specific content on it. Although this limits the analysis because there could be many more collaborations between organizations that were not considered in this study, it focuses the analysis on data that is already transparent and collaborations that are happening in the open.

Chapter 4

4 Conclusions and Future Research

4.1 Conclusion

As more people get digitally connected through the Internet, there is the potential to create innovations of public value through an open collaboration approach. These innovations have public value because everyone can be involved in making them, and the resulting product can be shared as open source software back to all citizens (Baldwin & Von Hippel, 2011). Government organizations are motivated to collaborate with each other because open collaboration can help them break down silos of knowledge, share code with others, and not have to recreate other people's work (Longo & Kelley, 2016; Mergel, 2015). Using online platforms such as GitHub can help them speed up their open government initiatives, improve data transparency, public participation, and crowdsource from all users on the platform. However, governments are challenged by the lack of policy covering what software is appropriate for them to use, and the training involved to get their staff to learn the technologies involved in open collaboration.

At least 50% of the government organizations on GitHub have actively used the platform within the past two years. However, they need to provide more metadata or contextual information about themselves in order to build trust with their citizens and encourage them to participate in their work. There is a collaborative relationship between most government organizations on GitHub, however the number of collaborations between a few organizations are much greater than most others, following a long tail distribution. This research confirms some of the findings from Mergel (2015) about organizations that were most collaborative because many of the organizations discovered to be collaborative in the USA are also collaborative with governments outside the USA. For example, @18F and @GSA, which are the chief open government organization from the USA had some of the highest centrality values, indicating that they are very central and influential in the social structure of these organizations.

This research adds to the body of knowledge about of the perspectives of GitHub use in the public administration in Canada that was started by Longo & Kelley (2016). Their work was looking at the use of GitHub in its earlier stages, which is right around the time of when the adoption of the platform by governments was at its peak in 2014. It also adds to the knowledge about how governments outside the USA view GitHub as a platform for open collaboration or to become an open government (Mergel, 2015). It also shows how the social structure of the organizations and users collaborating on GitHub can be analyzed using social network analysis to understand the relations between the organizations and their importance on each other (Otte & Rousseau, 2002). Knowing the network of the organizations and their relations, it could be possible for them to identify organizations they have not yet worked with and would be open to cooperation in the future, almost like a Facebook friendship network to see friends of friends.

As for the number of people and organizations connected to the Internet grows, the number of potential collaborators to make innovations of public value will keep increasing. There is a growing opportunity for government organizations to make the most of the open source trend of software and the open collaboration approach of platforms such as GitHub or GitLab. Using pre-existing open source software, or open collaboration to generate new government software could become a method to reduce the procurement cost of acquiring proprietary software. Open source software can be used by other government organizations without facing political challenges to request access to data or software, thus helping break down silos of knowledge and opening access to information that does not need to be proprietary.

This research shows that it is possible to use GitHub as a case study to understand how government organizations aim to become open governments, and how they participate in open collaboration. The research methods in this study demonstrate an empirical approach to quantitatively measure the user perspectives and collaborations on GitHub. The two research methods from Chapters 2 & 3 demonstrate two different approaches, but they both aim to empirically measure open government and open collaboration. Where the social network analysis zooms out to paint a large picture of the entire ecosystem of government collaborations, the SWOT analysis zooms in to focuses on the experiences of the individual users. Where the two research methods help each other is that the social network analysis techniques can be used to see how GitHub is currently being used, while the SWOT analysis and user interviews opens a discussion about how GitHub could be used in the future. For example, some of the themes identified in the SWOT analysis contain ideas about the future use of GitHub, such as the threat of privacy for Canadian data on GitHub after it was acquired by Microsoft, or the opportunity of the platform breaking down silos of knowledge, and encouraging communication of data and software between governments.

It may be possible that as there are new changes to the political climate, policies, and usage of open source/collaboration software for government work, some may choose to join the collaboration, and some may leave it if their priorities are not met. It is still uncertain if GitHub will be the definitive platform for open collaboration on government work because it faces competition from other software collaboration and hosting services such as GitLab, BitBucket, or SourceForge. However, work done so far in an open collaboration approach should be further supported to promote open government, open data, and open collaboration for all.

4.2 Future Research Directions

Although this research shows that a few government organizations are more collaborative than most others, it does not explain why exactly that is the case. There is an opportunity to analyze the open government policies of the organizations from countries that use GitHub and see if certain policies encourage or discourage open collaboration. It would also help to analyze the policies around the usage of open collaboration platforms like GitHub that are owned by American companies versus companies native to a country.

The GitHub user interview and SWOT analysis method could be repeated to study different GitHub user groups such as users from different countries, or different levels of government (federal, provincial, local) to identify how they use GitHub differently. The social network analysis methods could be repeated on a periodic basis (annual or monthly) in order to see changes overtime in the collaborative nature of government organizations. Additionally, collaborative users or organizations based on the social network could be identified as champions of open collaboration, and they could be contacted to provide their perspectives about the usage of GitHub, open collaboration, and government work being done on the platform.

In order to identify the motivations and challenges that the government organizations face when using GitHub, only Canadian workers were interviewed, specifically a handful from the Eastern regions of Canada. Future research could look to include a balanced proportion of interviewees from all parts of Canada, or at least all GitHub government organizations from Canada. Getting a broader range of perspectives could reveal the differences in how workers from different parts of Canada or different levels of government view the platform. The research could even include the perspectives from participants outside Canada, especially from organizations outside of North America, or from non-anglophone areas. Since GitHub is only available in English, it would be of interest to see how the usage of the platform differs for users who are non-anglophones or have English as their second language. There are various government organizations that use GitHub in Europe, South America, and Oceania that could provide insights from their experiences of how they use platform which could contrast the usage of the platform in the USA.

The network of government organizations can be further analyzed by trying to cluster the organizations based on their centrality measures in order to identify if there are any clusters of organizations who work together more closely than others. In a geographic context, government organizations are physically distant from each other, but in an online platform that connects a global pool of users, it is possible that users may group together based on their interests rather than their geographic location. The clustering of the users or organizations on collaborative platforms could be studied to see how it affects their volume and quality of their collaborations and resulting products.

4.3 Recommendation to GitHub

GitHub should encourage all government organizations on the GGC list to become a verified account on the platform. This would prevent any account that is not actually a government account from getting onto the GGC list and misleading citizens. For example, any user of GitHub could submit a pull request onto the GGC list to add an organization account that has the name of a real place or government, but with fake contact information and get approved to be added if the user responsible for vetting the list isn't aware of the fake copy. Even now, the GGC list contains a few organizations that just have their account on the list, but there is no information on the organization's profile page about contact info, public repositories, or public members in the group. GitHub should also improve their data quality requirements needed for an organization to be added onto the GGC list, specifically requiring them to provide all the contextual information about themselves such as: description, email, website, and geographic location. Government organizations on GitHub should try to make as many of their repositories to be public and have their organization members be publicly listed on the organization page. The overall transparency of information would encourage participation from other users on the platform who are interested in contributing in open source projects.

References

(2019, September 9). Retrieved from Internet Live Stats: https://www.internetlivestats.com/

(2019, September 25). Retrieved from Open Source Initiative: https://opensource.org/osd

Anderson, P. (2007). What is Web 2.0?: ideas, technologies and implications for education . *Bristol: JISC*, 1-64.

Anselin, L. (2012). Anselin, L. (2012). From SpaceStat to CyberGIS: Twenty years of spatial data analysis software. *International Regional Science Review*, 131-157.

Baldwin, C., & Von Hippel, E. (2011). Modeling a paradigm shift: From producer innovation to user and open collaborative innovation. *Organization Science*, 1399-1417.

Baytiyeh, H., & Pfaffman, J. (2010, November). Open source software: A community of altruists. *Computers in Human Behavior, 26*(6), 1345-1354. doi:https://doi.org/10.1016/j.chb.2010.04.008

Blischak, J. D., Davenport, E. R., & Wilson, G. (2016). A Quick Introduction to Version Control with Git and GitHub. *PLoS Computational Biology, 12(1)*, e1004668.

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence, 14(1)*, 75-90.

Brabham, D. C. (2010). *Crowdsourcing as a model for problem solving: leveraging the collective intelligence of online communities for public good.* The University of Utah.

Budhathoki, N. R., & Haythornthwaite, C. (2013). Motivation for open collaboration: Crowd and community models and the case of OpenStreetMap. *American Behavioral Scientist, 57(5)*, 548-575.

Carolan, B. V. (2013). *Social network analysis and education: Theory, methods & applications.* Sage Publications.

Criado, I. J., Sandoval-Almazan, R., & Gil-Garcia, R. J. (2013, October). Government innovation through social media. *Government Information Quarterly, 30*(4), 319-326.

Dabbish, L., Stuart, C., Tsay, J., & Herbsleb, J. (2012, February). Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository. *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 1277-1286.

Edelmann, N., Höchtl, J., & Sachs, M. (2010). Collaboration for open innovation processes in public administrations. *In Empowering open and collaborative governance*, 21-37.

Forte, A., & Lampe, C. (2013). Defining, understanding, and supporting open collaboration: Lessons from the literature. *American Behavioral Scientist, 57(5)*, 535-547.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.

Friedman, N. (2019, January 7). *New year, new GitHub: Announcing unlimited free private repos and unified Enterprise offering*. Retrieved from The GitHub Blog: https://github.blog/2019-01-07-new-year-new-github/

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature, 438*, 900-901.

GitHub Guides. (2019, September 27). Retrieved from Hello World: https://guides.github.com/activities/hello-world/

GitHub Organizations. (2019, 12 12). Retrieved from Github: https://github.com/search?q=type%3Aorganization&type=Users

GitHub Repositories. (2019, 12 12). Retrieved from GitHub: https://github.com/search?q=is:public

GitHub Users. (2019, 12 12). Retrieved from GitHub: https://github.com/search?q=type%3Auser&type=Users

Goodchild, M. F. (2007). Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. *IJSDIR, 2*, 24-32.

Goodchild, M., & Glennon, A. J. (2010, January). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth, 3(3)*, 231-241.

government.github.com. (2019, September 3). Retrieved from GitHub: https://github.com/github/government.github.com/pull/775

Gürel, E., & Tat, M. (2017). SWOT Analysis: A Theoretical Review. *Journal of International Social Research, 10(51)*.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenSteeetMap and Ordnance Survey datasets. *Environment and planning B: Planning and design, 37(4)*, 682-703.

Haklay, M. M., Basiouka, S., Antoniou, V., & Ather, A. (2010). How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The cartographic journal, 47(4)*, 315-322.

Haklay, M., & Budhathoki, N. (2010). OpenStreetMap—Overview and Motivational Factors. *Horizon Infrastructure Challenge Theme Day*.

Haklay, M., & Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, 12-18.

Halfaker, A., Geiger, S. R., Morgan, J. T., & Riedl, J. (2013). The Rise and Decline of Open Collaboration System: How Wikipedia's Reaction to Popularity is Causing Its Decline. *American Behavioral Scientist, 57(5)*, 664-688.

Harrison, T. M., Guerrero, S., Burke, G., Cook, M., Cresswell, A., Helbig, N., . . . Pardo, T. (2012). Open government and e-government: Democratic challenges from a public value prespective. *Information Polity, 17(2)*, 83-97.

Harrison, T., Pardo, T. A., & Cook, M. (2012). Creating Open Government Ecosystems: A Research and Development Agenda. *Future Internet, 4(4)*, 900-928.

Howe, J. (2006). *Crowdsourcing: A definition.*

(2015). *ICT Facts & Figures.* Geneva: Internet Telecommunications Union. Retrieved from https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, 258-268.

Johnson, P. A. (2019, January). Disintermediating Government: The role of Open Data and Smart Infrastructure. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Johnson, P. A., & Renee, S. E. (2012). Motivations driving government adoption of the Geoweb. *GeoJournal, 77(5)*, 667-680.

Kalin, I. (2014). Open Data Policy Improves Democracy. *Review of International Affairs, 34(1)*, 59-70.

Kelly, C., & Breinlinger, S. (1996). *The social psychology of collective action: Identity, injustice, and gender.* US: Taylor & Francis .

Kittuer, A., & Kraut, R. E. (2008, November). Harnessing the wisdom of crowds in wikipedia: quality through coordination. *In Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 37-46.

Knoke, D., & Kuklinski, J. H. (1982). Network Analysis.

Lampe, C., Wash, R., Velasquez, A., & Ozkaya, E. (2010, April). Motivations to Participate in Online Communities. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1927-1936.

Lee, G., & Kwak, Y. H. (2012). An Open Government Maturity Model for social media-based public engagement. *Government information quarterly, 29(4)*, 492-503.

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., . . . Wolff, S. (2009, October). A brief history of the internet. *ACM SIGCOMM Computer Communication Review, 39*(5), 22-31.

Leskovec, J., & Horvitz, E. (2008, April). Planetary-scale views on a large instant-messaging network. *Proceedings of the 17th international conference on World Wide Web*, 915-924.

Levine, S. S., & Prietula, M. J. (2014). Open Collaboration for Innovation: Principles and Performance. *Organization Science, 25(5)*, 1414-1433.

Liao, H.-T. (2009). Conflict and consensus in the Chinese version of Wikipedia. *IEEE Technology and Society Magazine*, 49-56.

Lima, A., Rossi, L., & Musolesi, M. (2014, May). Coding together at scale: GitHub as a collaborative social network. *In Eighth International AAAI Conference on Weblogs and Social Media*.

Longo, J., & Kelley, T. M. (2016). GitHub use in public administration in Canada: Early experience with a new collaboration tool. *Canadian Public Administration, 59(4)*, 598-623.

Maisonneuve, N., Stevens, M., & Ochab, B. (2010). Participatory noise pollution monitoring using mobile phones. *Information Polity, 15*, 51-71.

McDermott, P. (2010). Building Open Government. *Government Information Quarterly, 27(4)*, 401-413.

McDonald, N., & Goggins, S. (2013, April). Performance and participation in open source software on github. *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, 139-144.

Mergel, I. (2015). Open collaboration in the public sector: The case of social coding on GitHub. *Government Information Quarterly*(32), 464-472.

Morris, A. D. (1986). *The origins of the civil rights movement.* Simon and Schuster.

Mowery, D. C., & Simcoe, T. (2005). Public and Private Participation in the Development and Governance of the Internet. *The Limits of Market Organization*, 256-293.

Neset, T. S., Opach, T., Lion, P., Lilja, A., & Johansson, J. (2016). Map-based web tools supporting climate change adaptation. *The Professional Geographer*, 103-114.

OpenStreetMap Stats Report. (2019, 12 12). Retrieved from OpenStreetMap: https://wiki.openstreetmap.org/wiki/Stats

O'Reilly, T. (2005). Web 2.0: Compact Definition.

O'Reilly, T. (2007). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies, 1*, 17.

Osimo, D. (2008). Benchmarking eGovernment in the Web 2.0 era: what to measure, and how. *European Journal of ePractice, 4*, 37.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, 441-453.

Palomino, J., Muellerklein, O. C., & Kelly, M. (2017). A review of the emergent ecosystem of collaborative geospatial tools for addressing environmental challenges. *Computers, Environment and Urban Systems*, 79-92.

Peterson, K. (2013). The GitHub Open Source Development Process. Retrieved 8 18, 2019, from https://github.com/kevinpeterson/github-process-research

Pickton, D. W., & Wright, S. (1998). What's swot in strategic analysis? *Strategic change, 7(2)*, 101-109.

Quelin, B. V., Kiveleniece, I., & Lazzarini, S. (2017). Public-private collaboration, hybridity and social value: Towards new theoretical perspectives. *Journal of Management Studies, 54(6)*, 763-792.

Raymond, E. (1999). The Cathedral and the Bazaar. *Knowledge, Technology & Policy, 12(3)*, 23-49.

Resnick, P. (2001). Beyond Bowling Together: SocioTechnical Capital. *HCI in the New Millennium, 77*, 247-272.

Rusk, D., & Coady, Y. (2014, May). Location-Based Analysis of Developers and Technologies on GitHub. *International Conference on Advanced Information Networking and Applications Workshops*, 681-685.

Santana, A., & Wood, D. (2009). Transparency and social responsibility issues for Wikipedia. *Ethics and Information Technology*, 133-144.

Santos, D., & Cabral, L. M. (2009). GikiCLEF: Crosscultural issues in an international setting: asking non-English-centered questions to Wikipedia. *In Cross Language Evaluation Forum: Working notes for CLEF 2009*.

Schneider, J. (2013). Identifying, annotating, and filtering arguments and opinions in open collaboration systems. *Doctoral dissertation. Digital Enterprise Research Institute (DERI), National Unviersity of Ireland, Galway*.

Schrock, A. R. (2016, February). Civic hacking as data activism and advocacy: A history from publicity to open government data. *New Media & Society, 18(4)*, 581-599.

Scott, J. (1988). Social network analysis. *Sociology*, 109-127.

Sobkowicz, P., Kaschesky, M., & Bouchard, G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, 470-479.

Stein, K., Kremer, D., & Schlieder, C. (2015). Spatial Collaboration Networks of OpenStreetMap. *OpenStreetMap in GIScience*, 167-186.

Steiniger, S., & Geoffrey, H. J. (2009). Free and open source geographic information tools for landscape ecology. *Ecological Informatics*, 183-195.

Steinmacher, I., Conte, T. U., Gerosa, M. A., & Redmiles, D. F. (2015, February). Social Barriers Faced by Newcomers Placing Their First Contribution in Open Source Software Projects. *In Proceedings of the 18th ACM conference on Computer supported cooperative work & social computing*, 1379-1392.

Storey, M. A., Singer, L., Cleary, B., Figueira Filho, F., & Zagalsky, A. (2014, May). The (R)Evolution of Social Media in Software Engineering. *Proceedings of the on Future of Software Engineering*, 100-116.

Surian, D., Lo, D., & Lim, E.-P. (2010). Mining Collaboration Patterns from a Large Developer Network. *2010 17th Working Conference on Reverse Engineering*, 269-273.

Taeihagh, A. (2017). Taeihagh, A. (2017). Crowdsourcing: a new tool for policy-making? *Policy Sciences*, 629-647.

Taewoo, N. (2012). Citizens' attitudes toward open government and government 2.0. *International review of administrative sciences*, 346-368.

The White House. (2009, January 21). *Memorandum on Transparency and Open.* Retrieved from Federal Register: https://www.govinfo.gov/content/pkg/FR-2009-01-26/pdf/E9-1777.pdf

Thung, F., Bissyande, T. F., Lo, D., & Jiang, L. (2013, March). Network Structure of Social Coding in GitHub. *2013 17th European Conference on Software Maintenance and Reengineering*, 323-326.

Travers, J., & Milgram, S. (1977). An experimental study of the small world problem. *Social Networks*, 179-197.

Tredinnick, L. (2006). Web 2.0 and Business: A pointer to the intranets of the future? *Business information review, 23(4)*, 228-234.

Tsay, J., Dabbish, L., & Herbsleb, J. (2014, May). Influence of social and technical factors for evaluating contribution in GitHub. *Proceedings of the 36th international conference on Software engineering*, 356-366.

Ugander, J., Karrer, B., Backstorm, L., & Marlow, C. (2011). The anatomy of the Facebook social graph. *arXiv preprint arXiv:1111.4503*.

Van Zomeren, M., & Iyer, A. (2009). Introduction to the Social and Psychological Dynamics of Collective Action. *Journal of Social Issues, 65(4)*, 645-660.

Vasilescu, B., Posnett, D., Ray, B., van den Brand, M. G., Serebrenik, A., Devanbu, P., & Filkov, V. (2015, April). Gender and tenure diversity in GitHub teams. *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 3789-3798.

Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: An open data perspective. *Government Information Quarterly, 31(2)*, 278-290.

Vetro, A., Canova, L., Torchiano, M., Minotas, C. O., Lemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quartely*, 325-337.

Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M., Liu, Y., & Nyerges, T. L. (2013). CyberGIS software: a synthetic review and integration roadmap. *International Journal of Geographical Information Science*, 2122-2145.

Wanstrath, C. (2018, June 4). *A bright future for GitHub*. Retrieved from The GitHub Blog: https://github.blog/2018-06-04-github-microsoft/

Wellman, B., & Berkowitz, S. D. (1988). *Social structures: A network approach.* CUP Archive.

Wikipedia Statistics: All Languages. (2019, 1 31). Retrieved from Wikipedia: https://stats.wikimedia.org/EN/TablesWikipediaZZ.htm

Wirtz, B. W., & Birkmeyer, S. (2015). Open Government: Origin, Development, and Conceptual Prespectives. *International Journal of Public Administration, 38(5)*, 381-396.

Wright, S. C., Taylor, D. M., & Moghaddam, F. M. (1990). Responding to membership in a disadvantaged group: From acceptance to collective protest. *journal of Personality and Social Psychology, 58(6)*, 994-1003.

Yagoda, J. A. (2010). Seeing is Believing: The Detainee Abuse Photos and Open Government's Enduring Resistance to Their Release during an Age of Terror. *U. Fla. JL & Pub. Pol'y*, 273.

Yannis, C., & Euripidis, L. (2012). Participative public policy making through multiple social media platforms utilization. *International Journal of Electronic Government Research*, 78-97.

Yu, H., & Robinson, D. G. (2011). The New Ambiguity of Open Government. *UCLA L. Rev. Discourse, 178*, 59.

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

Zagalsky, A., Feliciano, J., Storey, M.-A., Zhao, Y., & Wang, W. (2015, February). The emergence of github as a collaborative platform for education. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1906-1917.

Appendix A: List of Participants

Participant A. (2018). Managerial work. Enterprise Strategic Planning, Office of the Chief Information Officer Treasury Board of Canada Secretariat, Government of Canada. Telephone interview on December 7, 2018

Participant B. (2018). Software developer work. Meteorological Service of Canada, Implementation of Operational Services Section (COMI), Environment and Climate Change Canada (ECCC), Government of Canada. Telephone interview on December 7, 2018

Participant C. (2018). Managerial work. Public Engagement and Marketing, Privy Council Office, Government of Canada. Telephone interview on December 19, 2018

Participant D. (2019). Managerial work. Office of the Chief Information Officer, Treasury Board of Canada Secretariat, Government of Canada. Telephone interview on January 3, 2019

Participant E. (2019). Managerial work. Digital Transformation Office, Treasury Board of Canada Secretariat, Government of Canada. Telephone interview on January 11, 2019

Participant F. (2019). Software developer work. Government of Ontario. Telephone interview on January 25, 2019

Participant G. (2019). Software developer work. Ontario Ministry of Training, Colleges and Universities, Government of Ontario. Telephone Interview on February 8, 2019

Participant H. (2019). Software developer work. Shared Services Canada (SSC), Government of Canada. Telephone Interview on April 4, 2019

Appendix B: Interview Script

Title of the study: GitHub Use for Government Related Work

Faculty Supervisor: Dr. Peter A. Johnson, PhD, Associate Professor, Department of Geography and Environmental Management, University of Waterloo, Canada. Phone: +1-(519)-888-4567 extension 33078. Email peter.johnson@uwaterloo.ca.

Student Investigator: Jaydeep Mistry, Department of Geography and Environmental Management, University of Waterloo, Canada. Email: jrmistry@edu.uwaterloo.ca.

Interview Questions:

1) What does your organization use GitHub for? Could it include:
    a. Collaborate on projects with users within your organization
    b. Hosting and sharing of code or data from projects
    c. Allowing citizens to view projects
    d. Allowing government workers in other organizations to view the projects
    e. Allowing users outside your organization to contribute to your projects
    f. Figuring out who worked on what part of a project
2) What types of GitHub repositories are owned by your organization? Could it include:
    a. Code base of websites
    b. Code base of mobile or web-based apps
    c. Code base of a software library or Application Programming Interface (API)
    d. Code of scripts for internal tasks
    e. Non-code related files
3) What are the benefits you have experienced through the use of GitHub?
4) What are some of the drawbacks or issues you have experienced through the use of GitHub?
5) How has the use of GitHub changed over the course of a project?

6) Has using GitHub made the pace of progress slower or faster? What are some of things that could explain the current pace of progress for your organization's projects that use GitHub?

7) Does your organization allow contributions from users outside your organization? Including other government workers, or citizens of the community? Why or why not? To what degree?

    a. If your organization allows contributions from anyone outside the organization, has this deterred or improved the pace of progress?

8) Has GitHub improved the transparency of your organization?

9) What would you recommend to another government organization if they are considering to use GitHub for project collaboration? Any lessons learned from your experiences?

Appendix C: Interview Recruitment Materials

Recruitment Email

Title of the study: GitHub Use for Government Related Work

Faculty Supervisor: Dr. Peter A. Johnson, PhD, Associate Professor, Department of Geography and Environmental Management, University of Waterloo, Canada. Phone: +1-(519)-888-4567 extension 33078. Email peter.johnson@uwaterloo.ca.

Student Investigator: Jaydeep Mistry, Department of Geography and Environmental Management, University of Waterloo, Canada. Email: jrmistry@edu.uwaterloo.ca.

Script Message:

Hello,

My name is Jaydeep Mistry and I am a MES student working under the supervision of Dr. Peter A. Johnson from the Department of Geography and Environmental Management at the University of Waterloo. I am contacting you to invite you to participate in a research study about analyzing the use of the collaboration tool called GitHub, by governments or for government related work.

Participation in this study involves you answering one online interview, taking approximately 20 to 30 minutes of your time. The purpose of the interview is to ask you about your experiences with using GitHub for government related work. It includes asking you about any issues or benefits you have faced from using GitHub for government, how projects have evolved over the use of GitHub, and how does your organization handle input from users outside your organization. Your contact information was obtained through the publicly available repository and user account information that can be searched through the GitHub's website and its REST API.

I would like to assure you that the study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee. However, the final decision about participation is yours.

If you are interested in participating, please contact me at jrmistry@edu.uwaterloo.ca and we can schedule a time for your participation. I will provide you with an Information Letter regarding the details of the study, as well as a formal Consent Letter needed to be completed before the interview.

Sincerely,

Jaydeep Mistry

Masters of Environmental Studies

Department of Geography and Environmental Management

University of Waterloo, Canada

UW Email: jrmistry@edu.uwaterloo.ca

Information Letter

Title of the study: GitHub Use for Government Related Work

Faculty Supervisor: Dr. Peter A. Johnson, PhD, Associate Professor, Department of Geography and Environmental Management, University of Waterloo, Canada. Phone: +1-(519)-888-4567 extension 33078. Email peter.johnson@uwaterloo.ca.

Student Investigator: Jaydeep Mistry, Department of Geography and Environmental Management, University of Waterloo, Canada. Email: jrmistry@edu.uwaterloo.ca.

To help you make informed decisions regarding your participation, this letter will explain what the study is about, the possible risks and benefits, and your rights as a research participant. If you do not understand something in the letter, please ask one of the investigators prior to consenting to the study. You will be provided with a copy of the information and consent form if you choose to patriciate in the study.

What is the study about?

You are invited to participate in a research study about analyzing the use of the collaboration tool called GitHub, by governments or for government related work. The objectives are to analyze: a) how governments are currently using GitHub, b) why governments are using GitHub, and c) how has the use of GitHub affected government work and their interaction with their community. This is important as there has been a growing global push to make governments more transparent and allow more citizen input, but there hasn't been any conclusive analysis to understand if the current approaches are beneficial to the government organizations or their community. This study will help answer the objectives of the research, as well as identify how governments can work better in the future through the use of collaborative tools. This study is being undertaken as a part of my (Jaydeep Mistry) MES research.

I. Your responsibilities as a participant

What does participation involve?

Participation in this study involves you answering one online interview, taking approximately 20 to 30 minutes of your time. The purpose of the interview is to ask you directly about your experiences with using GitHub for government related work. It includes asking you about any issues or benefits you have faced from using GitHub for government, how have projects evolved over the use of GitHub, and how does your organization handle input from users outside your organization.

The verbal interview will be audio recorded to ensure for accurate transcription and analysis. Overall, the questions of the interview will only ask about your experiences with using GitHub as a member of your organization, and will not ask about any information related to your personal life.

Please note that the interviews will be operated by Skype calls. There is always a risk your responses may be intercepted by a third party (e.g., other government agencies, hackers). University of Waterloo researchers will not collect or use internet protocol (IP) addresses or other information which could link your participation to your computer or electronic device without first informing you. If you prefer not to participate using this online method, please contact one of the researchers so you can participate using an alternative method such as a telephone call.

Who may participate in the study?

In order to participate in this study you need to be a member of an organization that has used GitHub for government related work, or that you have used GitHub yourself for projects related to government work, which could include but not limited to:

- writing code to collect, store, analyze, visualize data or information related to governments
- writing documentation to collect, synthesize, describe code or projects related to governments
- Interacting with any repositories owned by your organization that are hosted publicly on GitHub

II. Your rights as a participant

Is participation in the study voluntary?

Your participation in this study is voluntary. You may decide to leave the study at any time by communicating to the researcher that you want to discontinue your participation. During the interview if you ask to skip a question, end the interview or withdraw the participation, then the interviewer will do so as per the request.

You can request your data to be removed from the study until April 2019 as it is not possible to withdraw your data once papers and publications have been submitted to publishers. Your identity will be kept confidential and the information you provide will only be used for research purposes.

Will I receive anything for participating in the study?

You will not receive payment for your participation in the study.

What are the possible benefits of the study?

The study will benefit the academic community that research government by taking an in-depth look at the current state of how these government organizations across the world use GitHub, what benefits they seek through GitHub, and what issues they come across. Such knowledge can help government organizations improve themselves to become more transparent, build integrity within their community, as well as improve the pace of collaboration for future work. Any participant who chooses to complete the interview will have the options to be kept notified about the study findings and final report, at the end of their interview.

What are the risks associated with the study?

There are no known or anticipated risks associated with participation in this study.

Will my information be kept confidential?

The research team will know which data is from your participation, however your identity will be kept anonymous to anyone outside the research team. Even during the transcription of the audio recordings, I (Jaydeep Mistry) will personally transcribe the audio to the best of my ability, and no third-party service will be involved.

The researchers will keep a list of names and their correspond code in a list that is separate from the data of the study (audio records, transcriptions, etc.). Their names will be stored as is in a list, along with a key code, where the key code will be attached to the interview audio and transcriptions, but their real names will not be attached to it, in order to maintain confidentiality.

Your information will be securely stored on a password protected University of Waterloo computer on the university's main campus. Identifying information will be removed from the transcripts and the audio recordings will be deleted after I defend my thesis (expected to be April 2019). The transcripts and other electronic data will be retained for a minimum of 2 years, after which it will be destroyed. Only the research team will have access to the study data. If the data is being submitted for publication, then the identities of all participants will be made anonymous.

How will my data be shared?

Once all the data are collected and analyzed for this project, the researchers plan to share this information with the research community through seminars, conferences, presentation, and journal articles. If you are interested in receiving more information regarding the results of this study, or would like a summary of the results, you will have the option to provide your email address after the interview, and when the study is completed, anticipated by (May 2019), the researchers will send you the appropriate information.

III. Questions, comments, or concerns

Who is sponsoring/funding this study?

This study is funded by scholarship money received through the Social Sciences and Humanities Research Council of Canada (SSHRC).

Has the study received ethics clearance?

The study have been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE#40296). If you have any questions for the Committee then contact the Office of Research Ethics at +1-(519)-888-4567 extension 36005 or email ore-ceo@uwaterloo.ca.

Who should I contact if I have questions regarding my participation in the study?

If you have any questions about the interview, you can contact myself (Jaydeep Mistry) at the email jrmistry@edu.uwaterloo.ca; alternatively you could contact my supervisor Dr. Peter Johnson +1-(519)-888-4567 extension 33078 or email peter.johnson@uwaterloo.ca.

Thank you for your participation,

Jaydeep Mistry

Masters of Environmental Studies

Department of Geography and Environmental Management

University of Waterloo, Canada

UW Email: jrmistry@edu.uwaterloo.ca

Consent Form

Title of the study: GitHub Use for Government Related Work

Faculty Supervisor: Dr. Peter A. Johnson, PhD, Associate Professor, Department of Geography and Environmental Management, University of Waterloo, Canada. Phone: +1-(519)-888-4567 extension 33078. Email peter.johnson@uwaterloo.ca.

Student Investigator: Jaydeep Mistry, Department of Geography and Environmental Management, University of Waterloo, Canada. Email: jrmistry@edu.uwaterloo.ca.

Interview Consent Script:

Hello,

My name is Jaydeep Mistry. I am master's student, and my supervisor is Dr. Peter A. Johnson from the department of Geography and Environmental management, University of Waterloo, Canada. As a brief reminder, I will be asking you questions about your organization's use of GitHub for government related work.

The interview will take 30 minutes of your time. You can decide not to answer any particular question, or withdraw your participation at any time. I would like to assure you that your identity will be kept confidential. Any personal identifying information will not appear in any reports, papers, publications, or presentations resulting from this study.

Do you have any questions about the project?

Your verbal consent is going to be recorded in a consent log. If you do not have any further questions consent to the following items:

( ) Yes ( ) No → Do you agree to participate in this study?

( ) Yes ( ) No → Do you agree to the interview being audio recorded for accurate transcription and analysis?

( ) Yes ( ) No → Do you agree to the use of anonymous quotations in any thesis or publications that comes from this research study?

Thank you, we can start the interview now.

Feedback Letter

Dear ***[Name of Participant]***,

I would like to thank you for your participation in this study entitled "GitHub Use for Government Related Work". As a reminder, the purpose of the study was to analyze the use of the collaboration tool called GitHub, for governments or for government related work.

The data collected during the interviews will contribute to helping the academic community that researches governments by taking an in-depth look at the current state of how these government organizations across the world use GitHub, what benefits they seek through GitHub, and what issues they come across. Such knowledge can help government organizations improve themselves to become more transparent, build integrity within their community, as well as improve the pace of collaboration for future work.

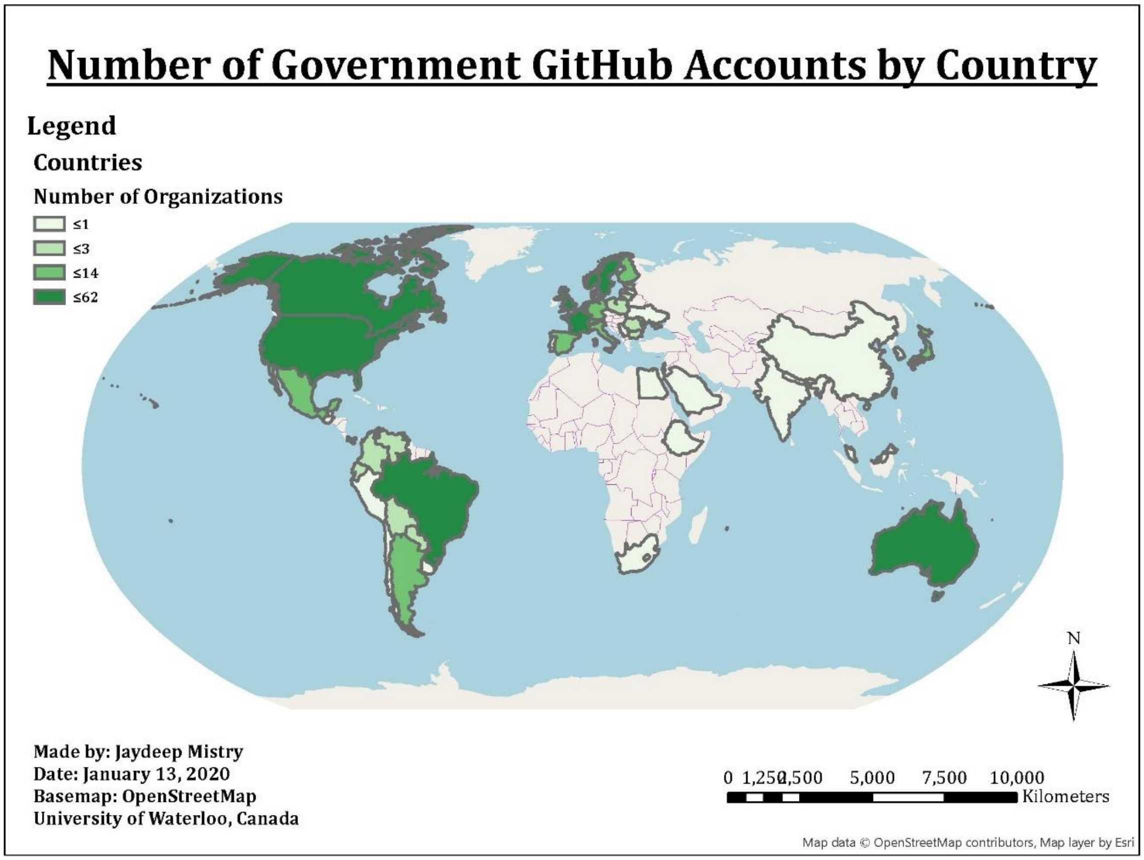
The study have been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE#40296). If you have any questions for the Committee then contact the Office of Research Ethics at +1-(519)-888-4567 extension 36005 or email ore-ceo@uwaterloo.ca.

If you have any questions about the interview, you can contact myself (Jaydeep Mistry) at +1-(226)-600-9560 or email jrmistry@edu.uwaterloo.ca; alternatively you could contact my supervisor Dr. Peter Johnson +1-(519)-888-4567 extension 33078 or email peter.johnson@uwaterloo.ca.

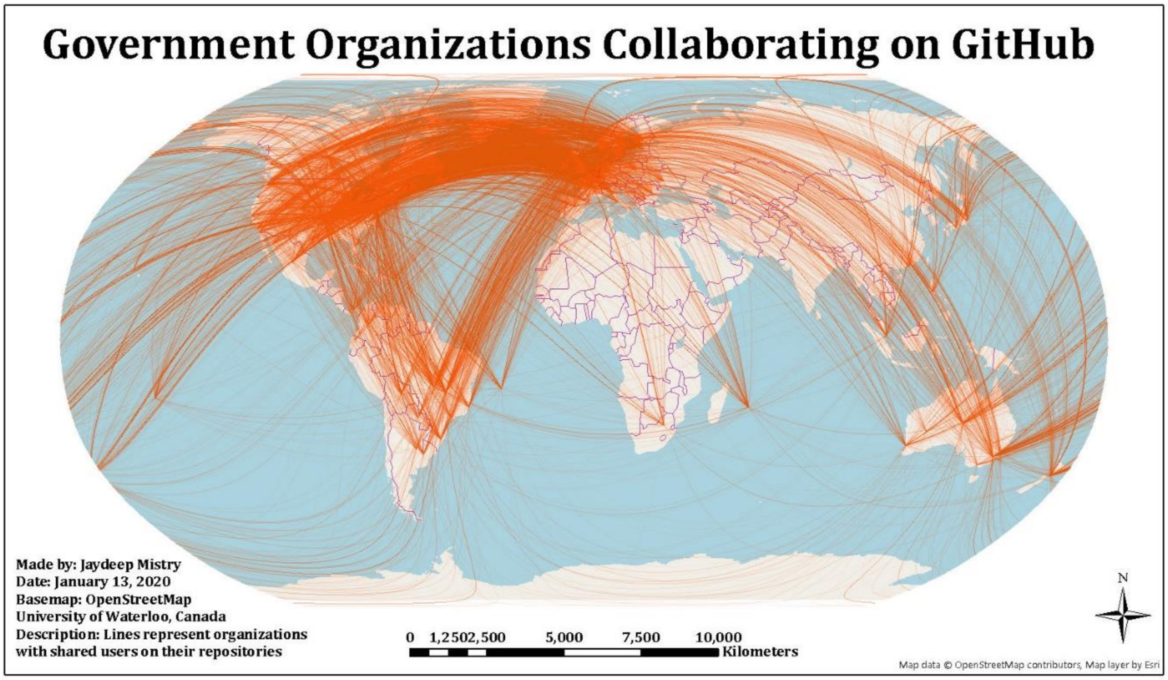
Please remember that any data pertaining to you as an individual participant will be kept confidential. Once all the data are collected and analyzed for this project, I plan to share this information with the research community through seminars, conferences, presentation, and journal articles. If you are interested in receiving more information regarding the results of this study, or would like a summary of the results, please provide your email address, and when the study is completed, anticipated by (May 2019), I will send you the information. In the meantime, if you have any questions about the study, please do not hesitate to contact me by email or telephone as noted in this letter.

Thank you for your participation,
Jaydeep Mistry
Masters of Environmental Studies
Department of Geography and Environmental Management
University of Waterloo, Canada
UW Email: jrmistry@edu.uwaterloo.ca

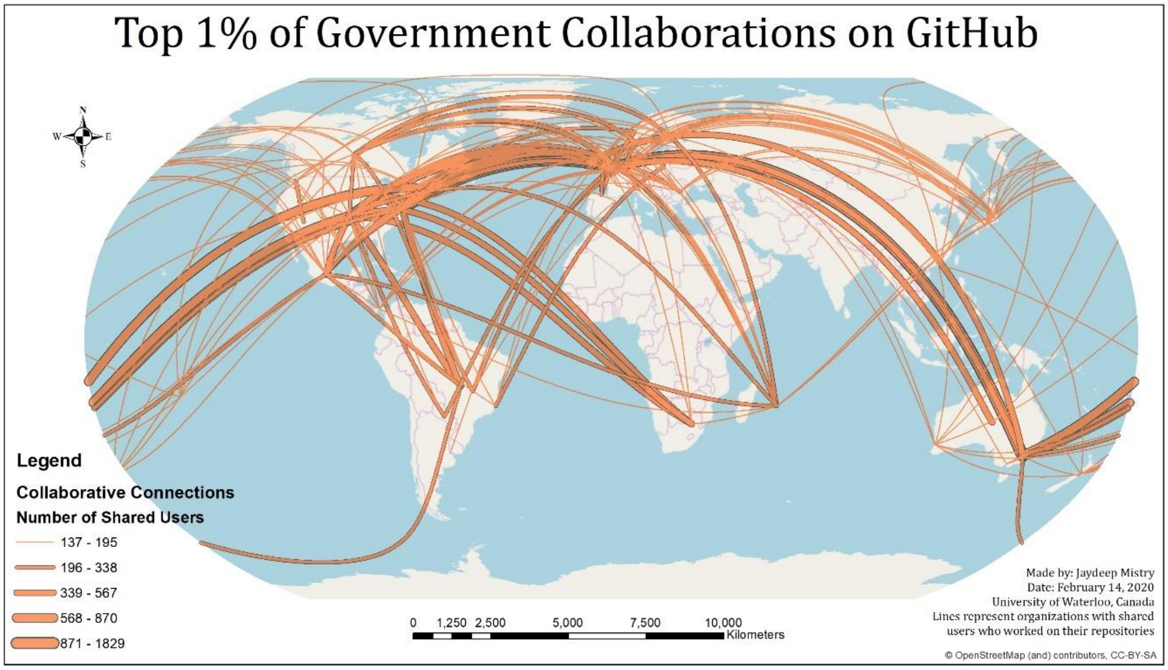
Appendix D: Maps of Government Collaborations



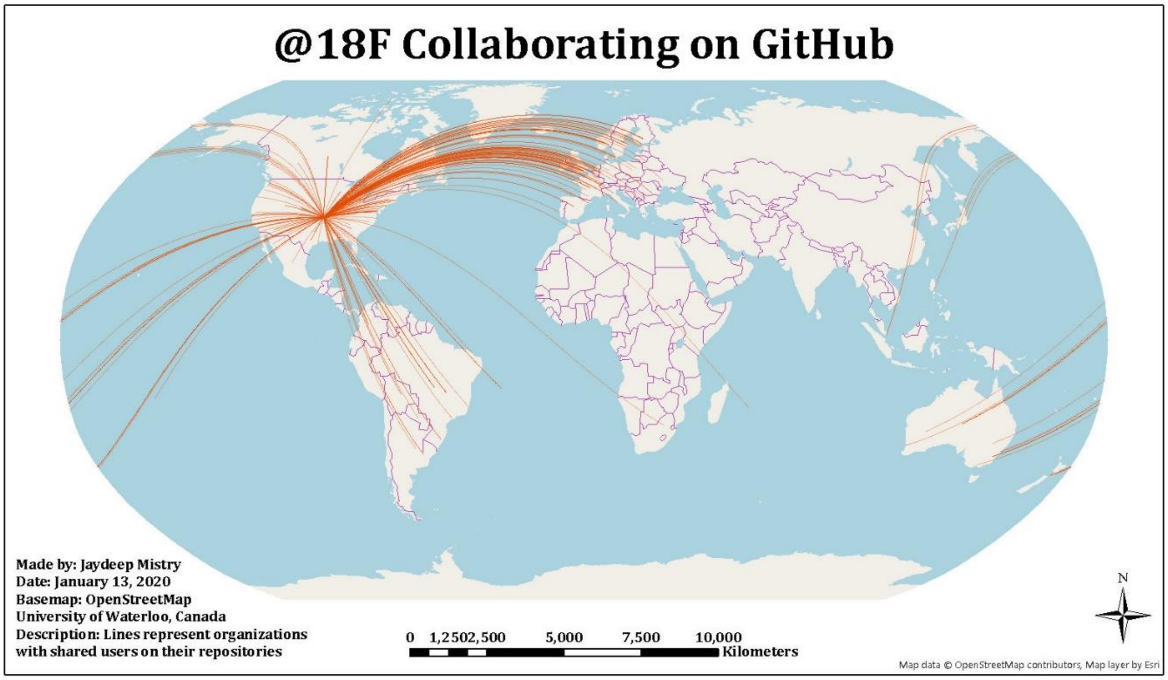
Map #1: Global map showing number of government GitHub accounts by country


Government Organizations Collaborating on GitHub
Made by: Jaydeep Mistry
Date: January 13, 2020
Basemap: OpenStreetMap
University of Waterloo, Canada
Description: Lines represent organizations with shared users on their repositories
0 1,2502,500 5,000 7,500 10,000
Kilometers
N
Map data © OpenStreetMap contributors, Map layer by Esri

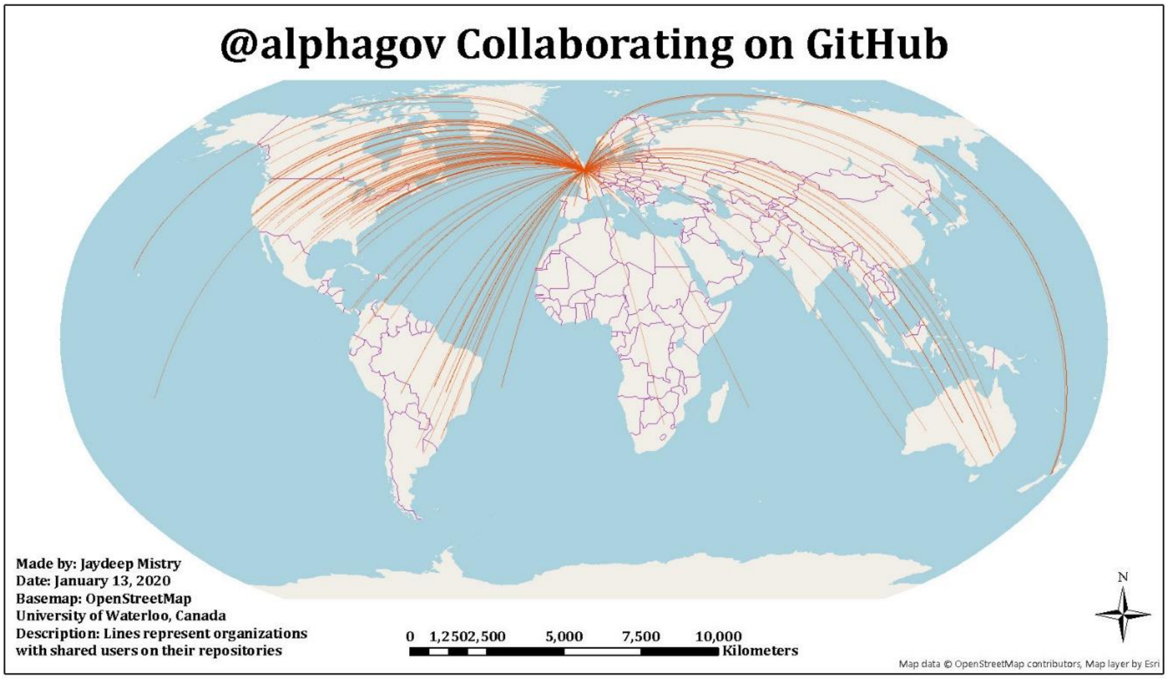
Map #2: Global map of all government organizations collaborating on GitHub


Top 1% of Government Collaborations on GitHub
N
W
E
S
Legend
Collaborative Connections
Number of Shared Users
137 - 195
196 - 338
339 - 567
568 - 870
871 - 1829
0 1,250 2,500 5,000 7,500 10,000
Kilometers
Made by: Jaydeep Mistry
Date: February 14, 2020
University of Waterloo, Canada
Lines represent organizations with shared users who worked on their repositories
© OpenStreetMap (and) contributors, CC-BY-SA

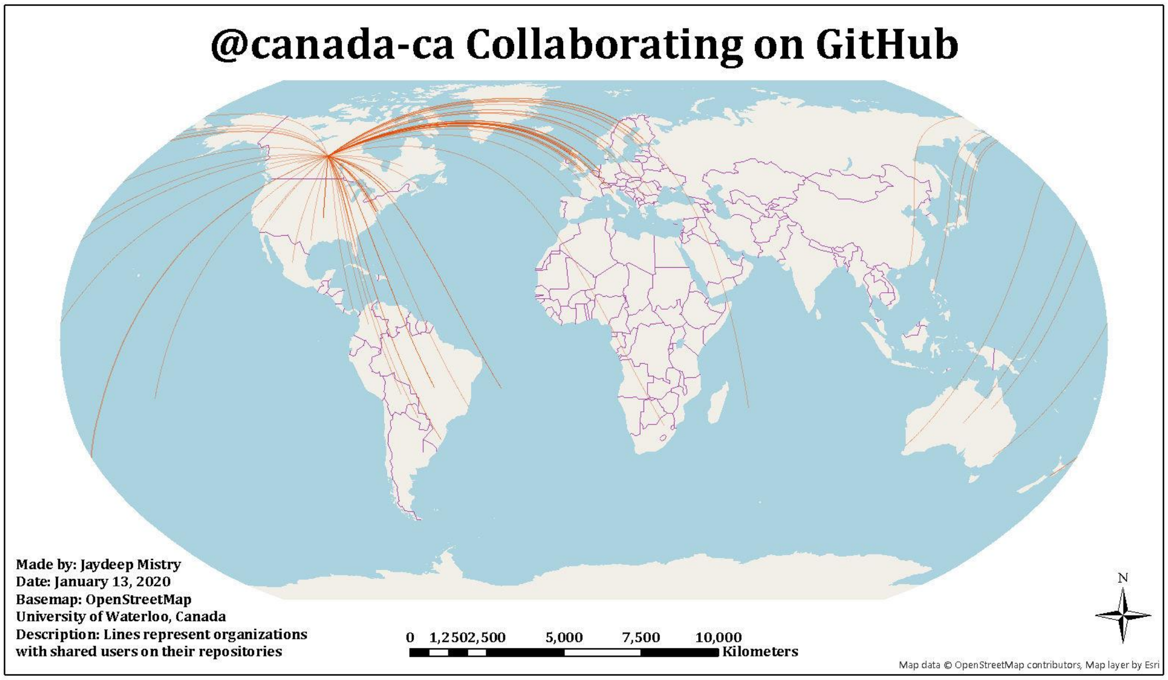
Map#3: Top 1% of the government organization collaborations on GitHub


@18F Collaborating on GitHub
Made by: Jaydeep Mistry
Date: January 13, 2020
Basemap: OpenStreetMap
University of Waterloo, Canada
Description: Lines represent organizations with shared users on their repositories
0 1,2502,500 5,000 7,500 10,000
Kilometers
N
Map data © OpenStreetMap contributors, Map layer by Esri

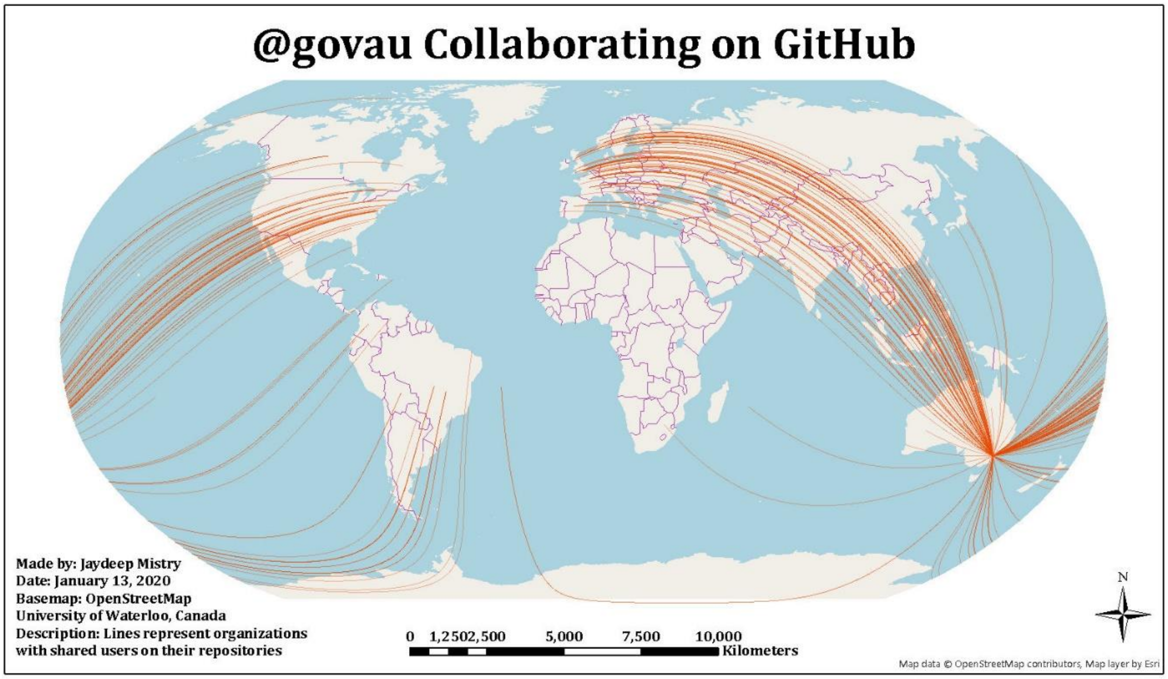
Map #4: Government organization @18F collaborating on GitHub


@alphagov Collaborating on GitHub
Made by: Jaydeep Mistry
Date: January 13, 2020
Basemap: OpenStreetMap
University of Waterloo, Canada
Description: Lines represent organizations with shared users on their repositories
0 1,2502,500 5,000 7,500 10,000
Kilometers
N
Map data © OpenStreetMap contributors, Map layer by Esri

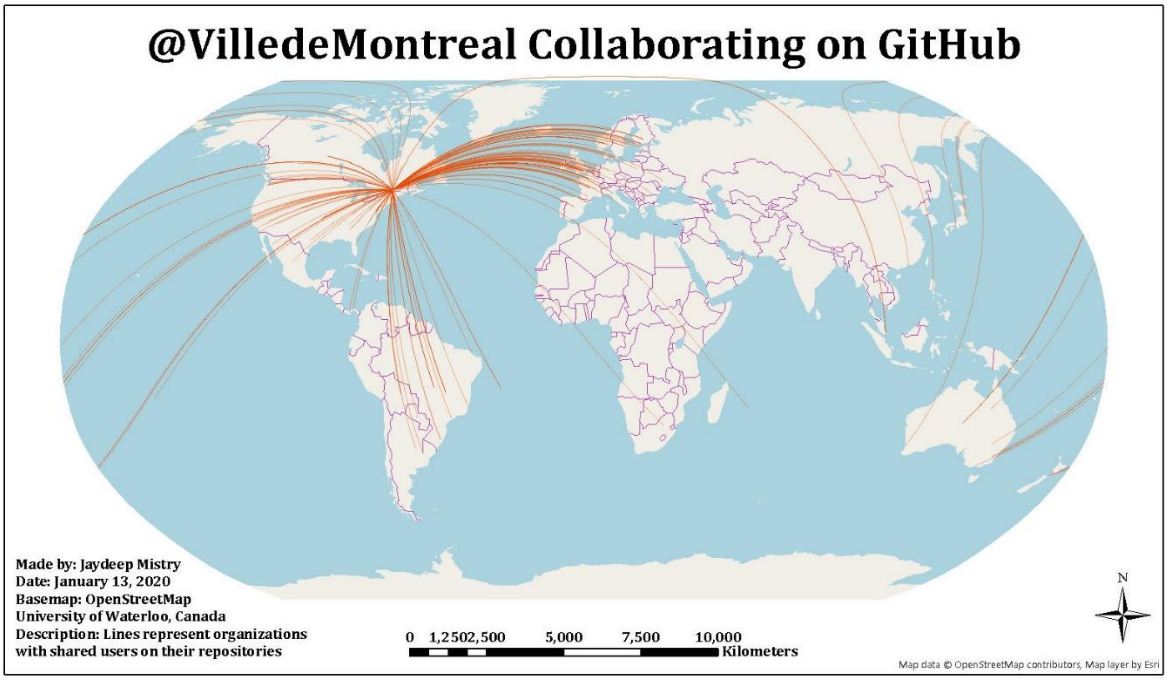
Map #5: Government organization @alphagov collaborating on GitHub


@canada-ca Collaborating on GitHub
Made by: Jaydeep Mistry
Date: January 13, 2020
Basemap: OpenStreetMap
University of Waterloo, Canada
Description: Lines represent organizations with shared users on their repositories
0 1,2502,500 5,000 7,500 10,000
Kilometers
N
Map data © OpenStreetMap contributors, Map layer by Esri

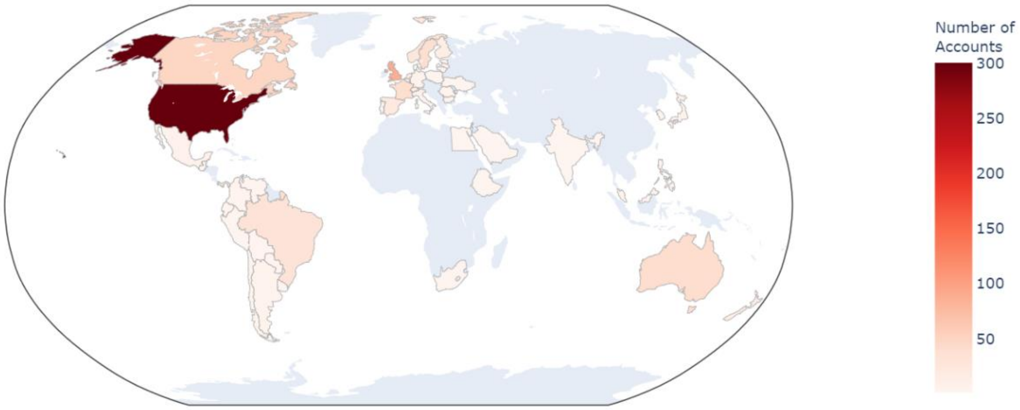
Map #6: Government organization @canada-ca collaborating on GitHub


@govau Collaborating on GitHub
Made by: Jaydeep Mistry
Date: January 13, 2020
Basemap: OpenStreetMap
University of Waterloo, Canada
Description: Lines represent organizations with shared users on their repositories
0 1,2502,500 5,000 7,500 10,000
Kilometers
N
Map data © OpenStreetMap contributors, Map layer by Esri

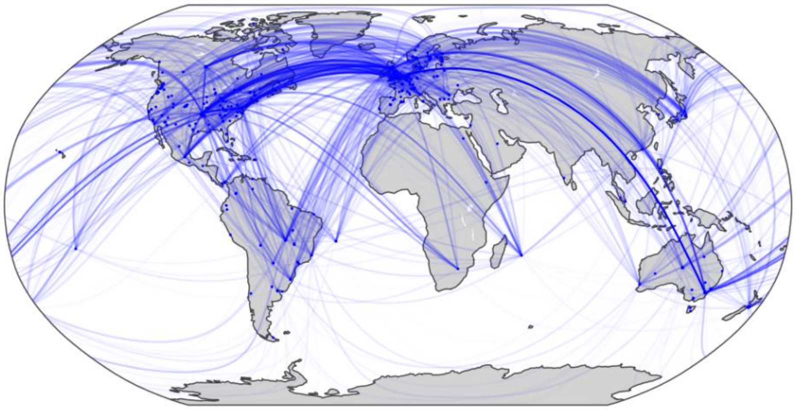
Map #7: Government organization @govau collaborating on GitHub


@VilledeMontreal Collaborating on GitHub
Made by: Jaydeep Mistry
Date: January 13, 2020
Basemap: OpenStreetMap
University of Waterloo, Canada
Description: Lines represent organizations with shared users on their repositories
0 1,2502,500 5,000 7,500 10,000
Kilometers
N
Map data © OpenStreetMap contributors, Map layer by Esri

Map #8: Government organization @VilledeMontreal collaborating on GitHub



Map #9: Simplified choropleth map of number of government GitHub accounts by country



Map #10: Simplified map of global network of government organizations collaborating on GitHub