Open vs. closed innovation: using online network data to measure innovation

Benjamin Snow and Oliver Bott

12 December 2014

Contents

Abstract: This research project examines the potential benefit of using open knowledge data in the form of collaborative online network data as an innovation indicator. By doing so, this work critically assesses current innovation indicators, namely patent data, in the hope of offering new alternatives for measuring and understanding innovation. The stated research question is: *To what extent can open innovation network data add to the measurement of innovation performance?* This study uses Github API follower data and patent data in 132 cities to compare the two, finding no correlation once controls are included in the model. However, several controls, including GDP, pollution and the relative importance of a city in its national context are found to be correlated with both patent data and open innovation data, suggesting that both types of innovation bear fruit in similar circumstances. We suggest that, while the study does not conclusively determine a future metric for measuring open innovation or comparing types of innovation, it does suggest a potential expansion of the current field of measuring innovation, which could potentially more accurately represent innovation in the open technological age.

---

1 Introduction

Policy makers worldwide have a profound interest in innovation for its significance for economic development and prosperity. Taylor (2004) views innovation as "the driving force behind modern economic growth, relative industrial power, and competitive advantage" (p.222). Numerous studies, for example the periodical Innovation Union Scoreboard[1] and the OECD Science, Technology and Innovation Scoreboard[2], have attempted to measure and compare innovation performance on the national level. However, until today most examinations of innovation have put their analytical emphasis on national level patent data, relying on this form of registering of proprietary data as a means of measuring innovation. The use of patent data has widely been criticized for its limited view of innovation, as patents are only filed as a means of protecting an idea for its exclusive commercial use. Many forms of innovation, with the key example being the opensource development of software, do not require or attempt government protection through patents, but would still in a digitalized society be considered as important forms of innovation.

Thus the academic literature until this point has hitherto ignored largely unexplored other, more open measures of innovation. The relatively recent emergence of network-based research systems offer new and potentially more instructive metrics by which to measure innovation, compared to protected closed knowledge. Since various scholars called for the continuous improvement of innovation measurement (see for example Freeman and Soete 2009), this work seeks to go beyond the widespread use of patent data to contribute to the refinement of innovation indicators, and the field as a whole.

In the following section, the current state of the field will be examined. Current best practice of defining and measuring innovation will be discussed before highlighting how new open innovation can improve the measurement of innovation. Further, other factors influencing innovation intensity will be discussed, which will inform our later analysis. In the third section, our methodology will be stated, including the process of data gathering, selection and analysis. This will be followed by our main section, the analysis, in which we use descriptive and inferential statistics to answer our research question. In a first step, we analyze the relationship between open and closed innovation in the form of collaborative online research platform use and

[1]For the latest edition see http://ec.europa.eu/enterprise/policies/innovation/policy/innovation-scoreboard/index_en.htm.
[2]For the latest edition see http://www.oecd.org/sti/scoreboard.htm.

patent intensity. In a second step, we look at the more fundamental drivers of innovation to establish whether there is an indirect relationship between the two variables of interest. Findings are further discussed and summarized in the concluding section. Overall, we find that, despite our exploratory research and various data limitations, there seems to be an indirect relationship between what we term open and closed innovation. The authors are confident that in the future similar research incorporating open innovation data will help improve the measurement of innovation and thus enhance our understanding of innovation.

2 State of the Field

2.1 Defining Innovation

Using a rather grand view in his understanding of innovation, Schumpeter (1942) sees innovation as "a process of industrial mutation, that incessantly revolutionizes the economic structure from within". In a more understated characterization, Smith (2005) defines innovation as "the creation of something qualitatively new, via processes of learning and knowledge building. It involves changing competences and capabilities, and producing qualitatively new performance outcomes" (Smith 2005, 149). While it is widely accepted that innovation can take many forms, e.g. product, process, marketing and proccess innovations, Frankelius (2009), in his extensive literature review of innovation studies, criticizes the widespread underlying assumption that innovation is limited to technological innovation. While accepting Frankelius (2009)'s critique of innovation as taking place outside of the technological realm, for the purpose of this study technological innovation, and specifically software innovation, will be the primary focus following relatively closely to Smith (2005)'s definition.

2.2 Measuring Innovation

The frequent technological focus when studying innovation can partly be explained by the difficulties associated with innovation's measurement. Smith (2005) notes the measurement challenge, as innovation is by definition a novelty and thus commensurability is a demanding task. For these reasons innovation has traditionally

though controversially been measured by looking either at its inputs, outputs, our throughputs. Attempting to measure innovation by inputs often focuses on resources, such as personnel and equipment allocated to research and development (R&D), which Freeman and Soete (2009) notes is often overestimating innovation in R&D by including the routine with the novel. Put another way, the use of research and development funding to assess innovation assumes innovation to takes place linearly with enough resources, as if the doubling of the number of Austrian patent office workers would have somehow resulted in two Albert Einstein's coming from their ranks, rather that just one. Freeman and Soete (2009) compares this to output oriented measures, which are often based on what he concludes are the already inadequate measures such as GDP. As Freeman and Soete (2009) suggests, GDP is an often cited imprecise statistical measure. However, building off of the first example, the use of output oriented measures the assumed result of innovation, economic growth, and not only assumes that the growth was based upon innovation, but seems to similarly assume that innovation creates value in a linear manner. A simple check of the wealth earned from various patents suggests this is not the case. Thus both input measures, such as R&D funding, and output measures, such as GDP, both either do not directly measure innovation, or do so in a manner so broad as to be unhelpful.

An indicator most often found in innovation research is patent data (see Taylor 2004). A patent is a "temporary legal monopoly granted by the government to an inventor for the commerical use of the invention [Taylor (2004), 229). A patent constitutes a property right awarded when an invention is shown to be non-trivial, useful, and novel (Taylor 2004, 230). Patents were first used to measure demand-side determinants of innovation, and have been used in the analysis of innovation activity for over three decades [Taylor (2004), 230). Taylor (2004) uses patent data taken from 1963 to 1999 in six different industries and their future citation levels and uses an Ordinary Least Squares (OLS) model to test what he terms the 'industry-innovation assumption'. The use of citations with patent data suggests a more nuanced examination of innovation using patent data than R&D fuding or GDP, as by using relative citation levels Taylor (2004) was able to weight the relative importance of a patent.

2.3 Limitations of Patent Data

Despite the usefulness of patent data, Taylor (2004) finds several limitations. In addition to the 'classification problem' related to assigning specific industries to patents, patents vary widely in significance, both technically and economically (Taylor 2004, 231). Most significantly for the purpose on this study, Taylor (2004) as well as Pakes and Griliches (1980) find that "patents are a flawed measure particularly since not all new innovations are patented and since patents differ greatly in their economic impact" (Taylor 2004, 378). Taylor (2004) to some degree is able to take into account the relative importance of different patents by taking into account their future citation level, a relatively good proxy for importance. Still he is less able to tackle the problem of new innovations which are not patented, with patent levels subsequently underrepresenting innovation. Thus, while for some considerable time patents have been considered to be the most effective proxy with which to measure innovation, they themselves ascribe to the notion that there is room for improvement in the study of innovation. This study stands as an attempt to further this field, to attempt to diliniate innovation which would not appear in patent data, but is instead based upon open network data.

2.4 Using Network Data as Innovation Indicator

Current developments in innovation research indicate that "characteristics that were important last century may well no longer be so relevant today and indeed may even be positively misleading" (Freeman and Soete 2009, 3). A shift away from the belief that innovation only occurs in professional R&D labs has occured, a change towards what Freeman and Soete (2009) calls "research without frontiers" (p.13). Even though networks and research collaborations become increasingly important, there have been relatively few studies focusing on network data (see Breschi and Malerba 2005). Even where research networks have been analyzed, the focus is too often on economically useful knowledge (see Acs, Anselin, and Varga 2002). Other studies focusing on research networks focus on other protected collaborative networks (see Ponds, Van Oort, and Frenken 2010). In an exception to this standard, Senghore et al. (2014) attempt to answer whether social network statistics act as indicators of innovation performance within a network, and which statistics could predict innovation performance. Using Gnyawali and Srivastava (2013)'s use models on cluster and network

effects to analyze multipartite social networks at mass collaboration events, gathering their data from NASA's International Space Apps Challenge. They use graph theory models constructed from affiliation networks finding (preliminarily) that distributions likely correlate to key aspects (Senghore et al. 2014).

Since Freeman and Soete (2009), among others, calls for the continuous improvement of innovation measurement, this work seeks to contribute to the refinement of innovation indicators. The purpose of this study is to explore the conceptual and statistical viability of a new metric by which we can measure innovation.

2.5 Fundamental Drivers of Innovation

While most previous studies of measuring innovation, as seen above, focus on the inputs (R&D) or outputs (GDP) of innovation, which this study attempts to reconsider, other research has attempted to define what actually drives or fosters innovation. While not the main focus of this study, which looks at measurement rather than impetus, an examination of what is considered plausible as a cause of innovation seems highly relevant.

When considering innovation drivers, a common culprit is creativity. While this may seem somewhat redundant, as innovation and creativity often in common langauge are used synonymously, for academic purposes innovation is the output of, and requires, creativity, but must be taken separately. However, the importance of creativity to innovation, and as a contributing predictor or cause of innovation has often been examined. Florida (2006), when discussing the creative economy, cites the importance of creativity not only in the science and technological fields, but the wider economy as a whole. He emphasizes the importance of schools and universities fostering creativity, and additionally notes the importance of geography and proximity to creative innovation, naming major global cities and imploring readers to move to one of these hubs, as the innovative as well as economic divide between these hubs and the rest of the world grows wider. To attract creative talent, deemend necessary for innovation, cities need to have certain characteristics being sought after by these creative groups (Florida 2006; Johnson 2014).

To study not only the direct relationship between patent and open innovation requires also taking into account other variables which either must be controlled for when evaluating both forms of innovation, or which could

explain one or another form of innovation in a specific locale, to explain what attributes of a location predict or explain innovation rather than simply measuring it. Many studies have attempted to evaluate different attributes of innovative regions, with policymakers wishing to find the elusive set of policies or conditions under which innovation flourishes.

Porter (1991) originally hypothesized that relationship to the environment, and specifically environmental regulation, could have a positive effect on innovation, as it forces the hand of domestic firms to be more innovative than their foreign counterparts, to remain competitive despite the increased environmental regulation. However, this initial hypothesis was not found in the evidence when tested by Jaffe and Palmer (1997), finding little evidence that industries become more inventive due to compliance costs. Wider examinations of the specific environmental conditions of a locale, its pollution level or amount of greenspace, for instance, affect people's level of innovation in these areas, or whether high mobility of labor allows innovators to selectively choose where they innovate. Along this line of thinking, Florida (2006) and Johnson (2014) mentioned previously in relation to creativity and innovation, suggest that, as creativity is a flow rather than a stock and requires an open system and longterm development, it can only occur in places where people can easily get to, lead the lives they want, and express themselves.

Past environmental conditions, labor conditions have long been studied as potentially relevant to the rate of innovation. Most often oversimplified as a cause and effect statement "Does technology create or destroy jobs" in which innovation in technology is the culbrit of economic disruptions. However, Pianta (2005) instead examines how innovation and employment effect each other when composition of skills and wages, and employment are taken into account, but often looks at innovation on the organizational, rather than the societal, level. This is why this study also takes into account variables that can explain innovation intensity in a given location, namely the GDP, total urban population, urban greenspace available, pollution levels and relative importance of a city in relation to its national context. Empirical findings in the literature suggest that these factors can affect the attraction and retaining of creative talent fundamental for any form of innovation activity.

2.6 Research Question

In light of the above mentioned state of innovation research we plan to examine the following research question:

To what extent can open innovation network data add to the measurement of innovation performance?

Exploiting technological advances related to the increasing use of the internet and open research platforms like GitHub[3], we plan to explore whether open knowledge networks can help refine currently limited innovation performance measurements.

3 Methodology

3.1 Data Gathering

To examine open network data against patent data, this study relies on two key data sources and uses the statistical tool *R* (R Core Team 2014) for the data analysis. We take city-level data of 132 cities overall, ranging from sixteen different OECD member countries (see Figure 1).

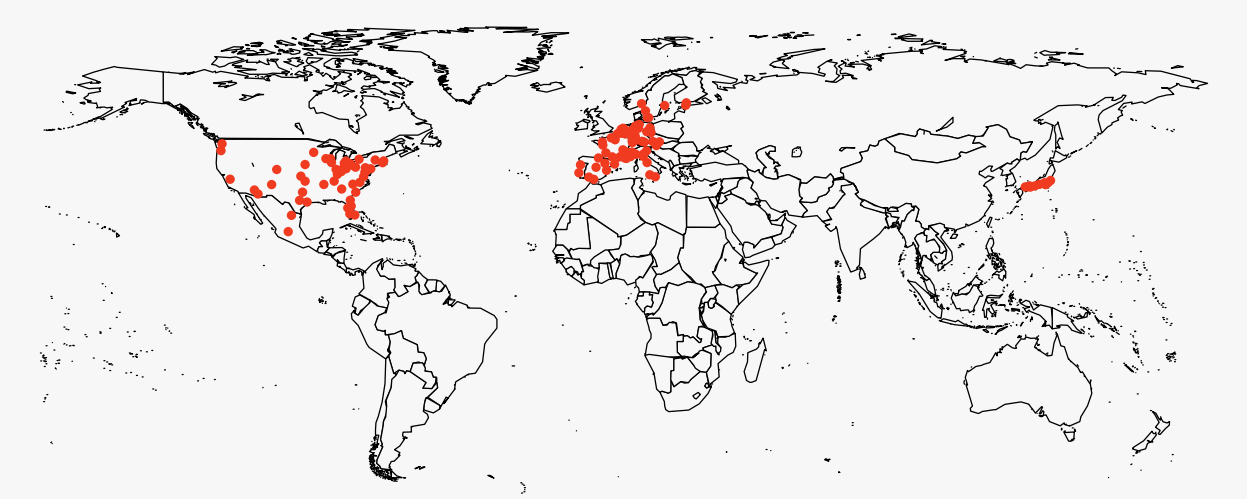

Figure 1: Locations of cities in the sample.

API network data

The first data set is obtained by using the Application Programming Interface (API) data for open networks.

[3]Online accessilbe on https://github.com/.

To examine open innovation, data is obtained from the the git repository web-based hosting service GitHub[4]. Its use of source code management makes it a commonly used software development collaboration tool. Since most of the repositories are openly accessible one can use API tools to track the popularity of contributors through a process called following. The *R* (R Core Team 2014) packages *httr* (Wickham 2014a), *dplyr* (Wickham and Francois 2014) and *rjson* (Couture-Beil 2014) allow for compiling data on the follower counts and locations associated with different users and online repositories.

Closed innovation OECD patent data

For closed innovation we use city-level patent data, taken from the Organization for Economic Co-operation and Development[5]. Patent Cooperation Treaty (PCT) patent data are used to track internationally patented inventions. The *R* (R Core Team 2014) package *rsdmx* (Blondel 2014) is necessary for obtaining the OECD dataset. As we are interested in patent data, we work with data indicating the PCT patent applications per 10,000 inhabitants. From the same database, we also use GDP per capita data and environmental data, as other variables thought potentially relevant in explaining differences in innovation.

To allow for reproducibility of this research, the code for gathering and cleaning the data is stored in a separate .R file and can be accessed here. Packages to clean, analyze and visualize the data include *R-car* (Fox and Weisberg 2011), *R-ggmap* (Kahle and Wickham 2013), *ggplot2* (Wickham and Chang 2014), *maps* (Brownrigg 2014), *maptools* (Bivand and Lewin-Koh 2014), *Rcpp* (Eddelbuettel and Francois 2014), *rCharts* (Vaidyanathan 2013), *RCurl* (Temple Lang 2014), *repmis* (Gandrud 2014), *reshape2* (Wickham 2014b) and *stargazer* (Hlavac 2014).

3.2 Data Sources

As is depicted in Table 1, the analysis is based on cross-sectional data with varying time frames. There are several limitations to the data used in this study. First, the time discrepency between different aspects of the data used, which range from 2005 to 2014, reflect an obvious data comparability constraint. The GitHub user data is taken from December 2014. Another limitation is that the study uses only cross-sectional, as opposed

[4]Online accessilbe on https://github.com/.
[5]Online accessible on http://stats.oecd.org.

to panel-data, meaning that trends over time are not observed, which further limits the internal and external validity of our findings. For data availability and access reasons, there are some prominent innovation hubs excluded, including San Fransisco and New York. Any found significance will need to take this into account. Perhaps most significantly, when comparing the two measures of innovation, it should be noted that patent data reflects innovation across all types of sectors, whereas Github data mainly reflects innovation within the software technology domain.

Other limitations are related largely to the nature or state of the online collaboration platform GitHub itself. First, GitHub diffusion, as a relatively new software offering, is growing but still not very high, and thus cannot be said to be fully representative. Additionally, GitHub operates in a limited number of languages (both computer languages and human spoken languages) and both human language and computer programming language adoption and preferences vary widely geographically, making comparability of innovation across regions difficult. Lastly, this study relies on users stating their place of residence, a piece of information which GitHub does not require of users, and which is not universally done. This study is only able to gather geographic information and follower numbers on users who voluntarily (and correctly) offer this information.

However, in defense of the dataset employed - with regard to the time discrepancy issue, many of the variables which are from different times would likely have limited change in a decade, for instance the available urban greenspace is not an aspect of a city which would change drastically in a few years or a decade. In regards to the exclusion of a few major tech spaces, whose non-inclusion limits the dataset, the use of over 130 cities should allow trends to be seen despite imperfect data. While the exclusion of some important innovation hubs like New York and San Fransisco should be taken into account, it seems implicitly that the studies finding would actually serve to provide better more general trends and relationships, by excluding tech havens and rather looking more broadly at the pace of innovation in most of the world.

Table 1: Data sources and explanations.

Variables	Explanation	Year	Source
Patents	PCT patents per 10,000 population	2008	OECD

Variables	Explanation	Year	Source
GDP	GDP per capita	2008	OECD
Population	Total urban population	2008	OECD
Greenspace	Green area per capita in square metres	2008	OECD
Employment	Employment of metropolitan area as % of national value	2008	OECD
Pollution	Annual average of pop exposure to air pollution PM2,5 in µg/m³	2005	OECD
Users	GitHub users with >20 followers per 10,000 population	2014	GitHub

3.3 Data Selection

Several potential explanatory variables are collected besides the patent and GitHub data. These variables were selected as they were thought to potentially show cause for why innovation, be it open or closed, occurs in a certain city, but needed to be variables that would not introduce endogeneity to the model.

Greenspace: The Greenspace indicator is deemed potentially relevant in that with a choice of city to innovate in (assuming some level of geographic labor flexibility) there might be a recreational value necessary for attracting talent. Put another way, green cities could attract innovators (see Florida 2006; Johnson 2014).

Pollution: The Pollution indicator is taken both as a broad proxy for the urban environmental conditions and also the level of industrialization (leaving aside a discussion of to what degree pollution is from industry vs cars), that it seemed worth exploring whether a certain level of pollution discouraged talent attraction of innovators on the city-level.

Employment: The Employment indicator is taken largely as an indication of that city´s significance within its national context. Understanding whether a city would likely be viewed as the most prominent or significant, and whether this effects innovation, or whether innovation takes place in smaller provincial cities, is worther understanding. Additionally, seeing if the type of innovation (open vs. closed) depends on the significance of the city is viewed as relevant.

GDP: A GDP indicator explores whether the size of the economy, or wealth generally, is related to innovation

on the city-level, and if it is indicator of one type of innovation over another.

Population: A Population variable explores whether there is a necessary city size threshold which corresponds to innovation, and also is taken for controlling for across cities, to find patent data or GitHub data per a number of people in a city.

GitHub user follower counts: The number of GitHub users with a certain follower count acts as our main variable of interest. To determine which follower range can be deemed as an indicator for innovative activities, we compile follower data in various categories ranging from users with 1-9, 10-19 up to 90-99 and over 100 followers (per 10,000 population). We choose GitHub users with more than 20 followers as our main indicator of high collaboration and innovative activity, as the threshold of 20 represents a drastic decline in user counts as can be seen in Figure 2.

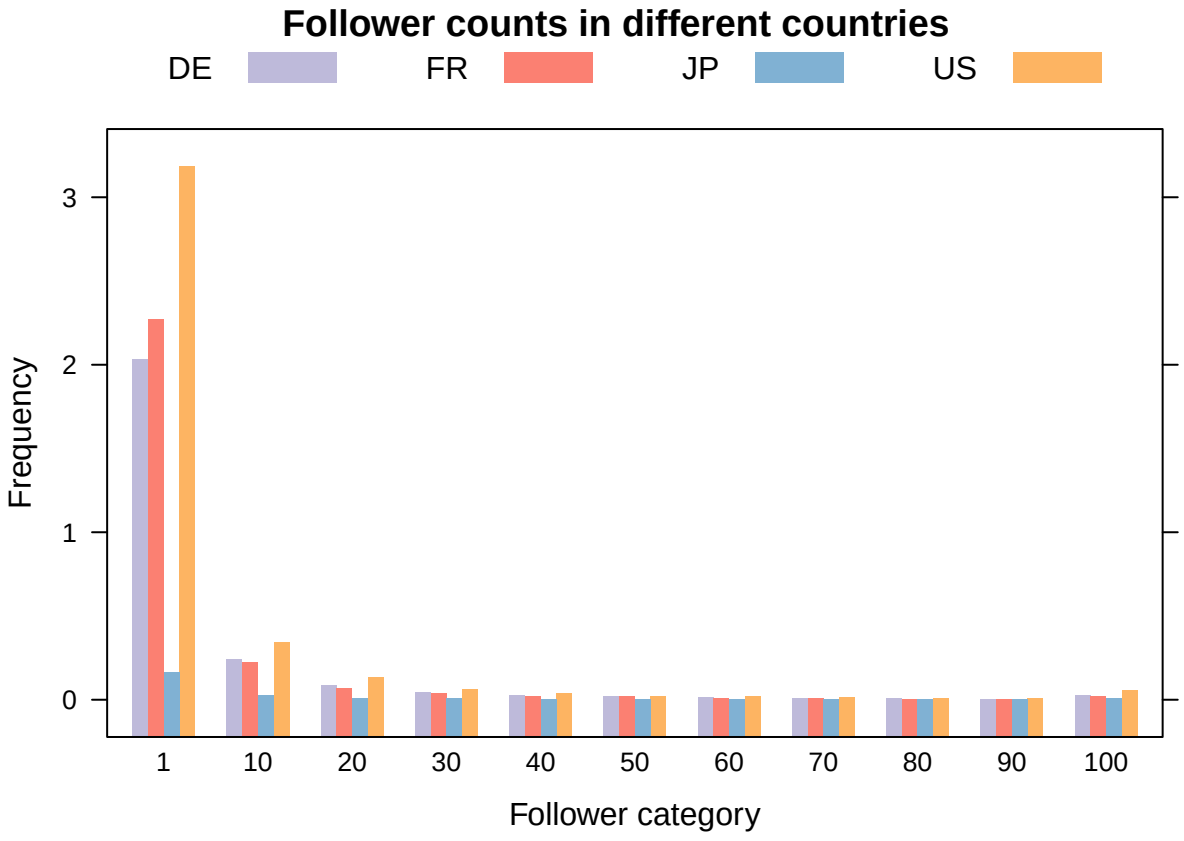

Figure 2: GitHub user's follower counts in various categories for cities in Germany, France, Japan and the United States.

4 Analysis

4.1 Descriptive Statistics

The summary statistics in Table 2 show wide ranging distributions of the observations in the data set. The *car* package (Fox and Weisberg 2011) is used to examine the relationship, distribution, and normality of all variables included in the model, to understand which regression model would be most appropriate. The distribution of many variables are highly skewed (see Figure 3). The GitHub based variable 'Users', our main variable of interest, is skewed to the right, as do nearly all of the observed variables, excluding Pollution, GDP and Patents, which come closer to a normal distribution. It seems as if already in the scatterplot a weak correlation between Patents and Users can be observed. To normalize for the skewed distributions, the log of the variables is deemed necessary to increase the explanatory power of our inferential statistics.

Table 2: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Max
Patents	132	1.69	1.39	0.06	6.86
GDP	132	36,308.69	8,575.53	5,133.12	61,804.14
Population	132	2,062,713.00	3,565,884.00	444,432	34,482,742
Greenspace	132	594.05	936.45	1.13	5,081.25
Employment	132	4.20	7.68	0.15	39.39
Pollution	132	16.21	5.09	5.85	31.44
>0 Followers	132	2.82	3.02	0.00	17.67
1-9 Followers	132	2.32	2.43	0.00	14.34
10-19 Followers	132	0.25	0.29	0.00	1.57
20-29 Followers	132	0.09	0.12	0.00	0.64
30-39 Followers	132	0.05	0.06	0.00	0.37
40-49 Followers	132	0.03	0.04	0.00	0.24
50-59 Followers	132	0.02	0.02	0.00	0.12
60-69 Followers	132	0.01	0.02	0.00	0.11
70-79 Followers	132	0.01	0.01	0.00	0.07
80-89 Followers	132	0.01	0.01	0.00	0.06
90-99 Followers	132	0.004	0.01	0.00	0.05
>100 Followers	132	0.03	0.05	0.00	0.32
>20 Followers	132	0.24	0.33	0.00	1.80

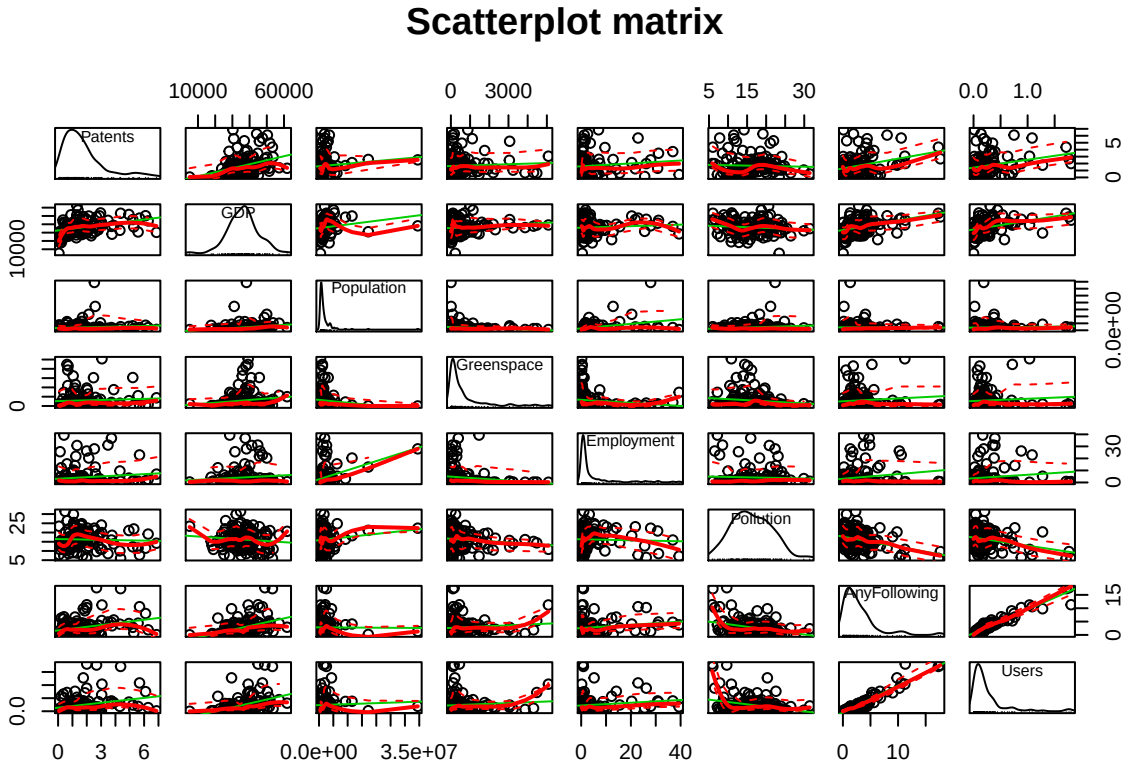

Figure 3: Scatterplot matrix of selected variables.

4.2 Inferential Statistics

To determine which regression model to choose for the inferential statistics, we use a residual plot. The residual plot between Patents and Users (see Figure 4) depicts a relatively random pattern, which indicates that a linear regression model provides a decent fit to the inferential statistics of the data set.

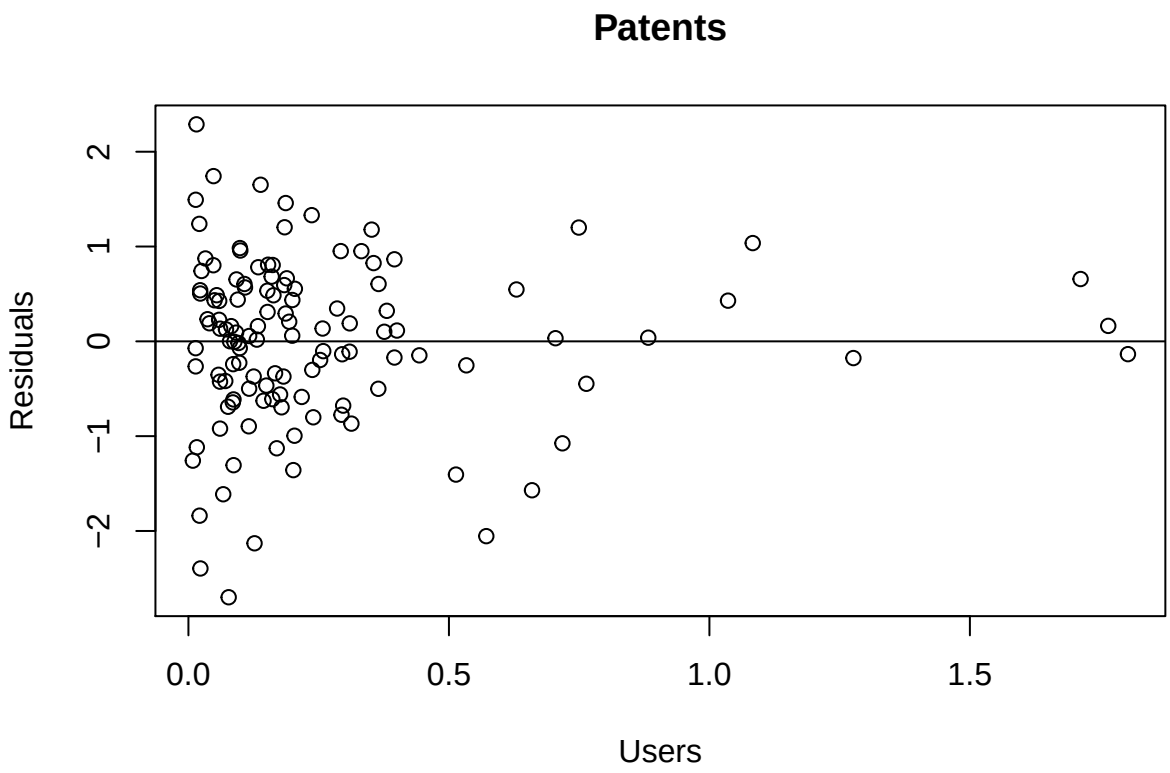

Figure 4: Residual plot for patents and users.

This study uses an ordinary least squares (OLS) model to examine the relationship between patent and highly followed open data sources. The first regression model includes the open innovation indicator U as the GitHub users with more than 20 followers expected in a given city i, and can be viewed as

$$logU_i = \beta_0 + \beta_1 logP_i + \beta_2 logGDP_i + \beta_3 logPop_i + \beta_4 logG_i + \beta_4 logE_i + \beta_4 logPoi_i + \epsilon_i$$

As can be seen in Table 3, the variable User is strongly positively correlated with the Patent variable (at

Table 3: Regression estimates of GitHub user counts

	Dependent variable:		
	log(Users)		
	(1)	(2)	(3)
Patents	0.37*** (0.11)	0.11 (0.11)	0.11 (0.13)
GDP		2.10*** (0.49)	1.54*** (0.58)
Population		−0.08 (0.14)	−0.22 (0.20)
Greenspace		0.05 (0.06)	−0.01 (0.06)
Employment		0.09 (0.07)	0.34* (0.17)
Pollution		−0.74*** (0.25)	−0.60** (0.28)
US			1.01* (0.58)
DE			0.56 (0.39)
FR			0.65* (0.38)
JP			−0.48 (0.57)
(Intercept)	−1.99*** (0.10)	−21.10*** (4.71)	−13.99** (6.29)
Observations	123	123	123
R^2	0.09	0.32	0.36
Adjusted R^2	0.08	0.28	0.30
Residual Std. Error	1.06 (df = 121)	0.93 (df = 116)	0.92 (df = 112)
F Statistic	12.30*** (df = 1; 121)	9.01*** (df = 6; 116)	6.31*** (df = 10; 112)
Note:			*p<0.1; **p<0.05; ***p<0.01

a significance level of p<0.01). Without including controls, a 1% increase in patents corresponds to a 0.37% increase in GitHub users with over 20 followers. However, the correlation becomes much smaller and insignificant when adding control variables. The GDP (at a significance level of p<0.01) and relative significance of a city in its national context (at a significance level of p<0.1), represented by the Employment variable, are positively correlated with GitHub users, while the pollution levels negatively correlate with our User variable (at a significance level of p<0.05). The urban greenspace and total population size of the city do not seem to have a correlation with GitHub users.

When adding country dummies, it becomes clear that both the United States and France seem to have more GitHub users than other countries in the sample. No effect can be seen for German and Japanese cities. The dummies must be interpreted with care, however, since the sample sizes for cities in these countries are very small, explaining the relatively large standard error. The adjusted R squared value indicates that, for our final specification of the first model, about 30% of the variation in users in our sample is explained through the model. The weak relationship between GitHub users and patents also becomes evident in the scatterplot in Figure 5, where we can witness a wide dispersion between the two variables. The distribution of the data

points indicates that some cities have relatively many GitHub users with more than 20 followers while some cities have close to zero users per 10,000 population.

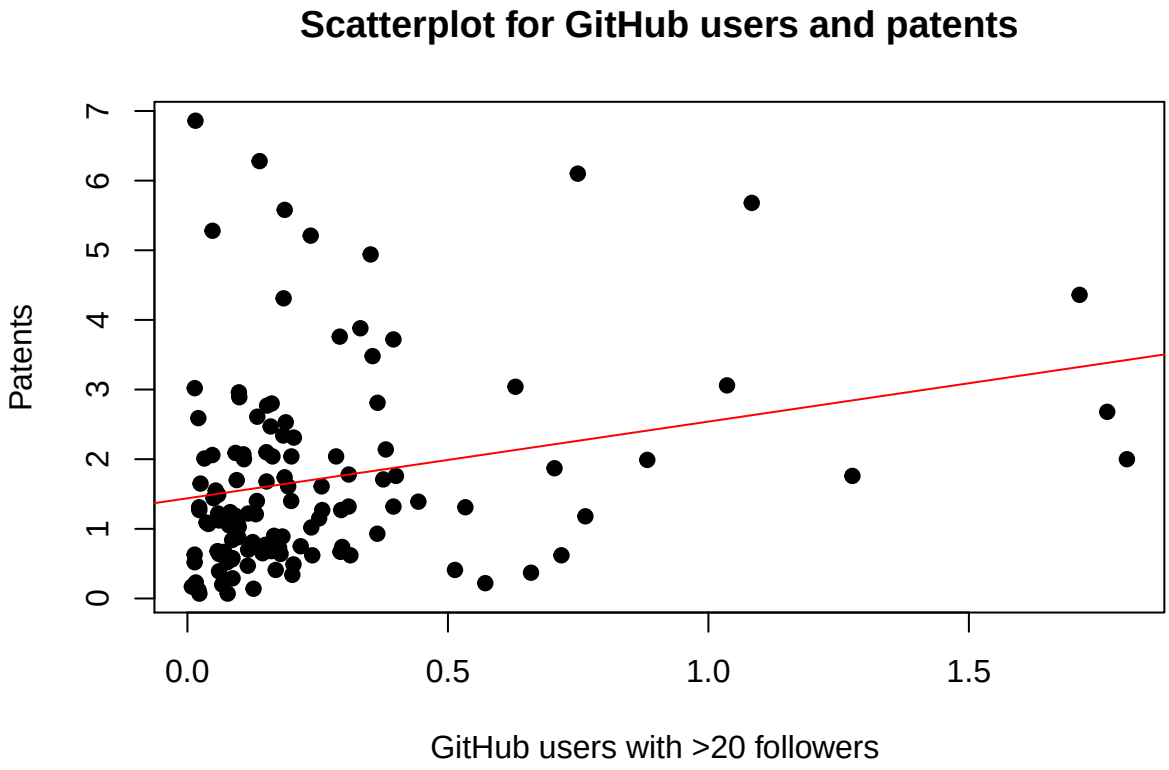

Figure 5: Scatterplot for GitHub users and patents with line of best fit.

Although there does not seem to be a direct relationship between patent and GitHub user data, further inferential statistical analysis attempts to find the common predictor or cause of innovation in both patent and open data. Our second model hence uses patents as the dependent variable P and is expressed below using similar notation and logic as stated above:

$$logP_i = \beta_0 + \beta_1 logU_i + \beta_2 logGDP_i + \beta_3 logPop_i + \beta_4 logG_i + \beta_5 logE_i + \beta_6 logPol_i + \epsilon_i$$

Similar to the first model, the regression output for our second model (see Table 4) demonstrates a positive relationship between patent data and GitHub users (at a significance level of $p<0.01$), but only when no controls are added. In the full model specification, again GDP and relative importance of a city in its

Table 4: Regression estimates of Patent Activity

	Dependent variable:		
	log(Patents)		
	(1)	(2)	(3)
Over 20 Followers	0.25^{***} (0.07)	0.08 (0.08)	0.08 (0.08)
GDP		1.90^{***} (0.40)	1.90^{***} (0.40)
Population		-0.11 (0.11)	-0.11 (0.11)
Greenspace		0.05 (0.05)	0.05 (0.05)
Employment		0.11^{*} (0.06)	0.11^{*} (0.06)
Pollution		0.45^{**} (0.21)	0.45^{**} (0.21)
(Intercept)	0.68^{***} (0.16)	-19.58^{***} (3.82)	-19.58^{***} (3.82)
Observations	123	123	123
R^2	0.09	0.31	0.31
Adjusted R^2	0.08	0.27	0.27
Residual Std. Error	0.87 (df = 121)	0.78 (df = 116)	0.78 (df = 116)
F Statistic	12.30^{***} (df = 1; 121)	8.61^{***} (df = 6; 116)	8.61^{***} (df = 6; 116)
Note:			*p<0.1; **p<0.05; ***p<0.01

national context positively correlate with Patents. Interestingly, however, the pollution level is also positively correlated to Patents (at a significance level of p<0.05), which is counterintuitive to the findings of Jaffe and Palmer (1997) and Florida (2006). The adjusted R squared value indicates that, for our final specification of the second model, about 27% of the variation in patents in our sample is explained through the model. Our analysis does not allow any statements as to the causal relationship of the variables in question, as we cannot eliminate potential endogeneity issues, e.g. through omitted variable bias, reversed causation or selection issues. This is why in our analysis we only speak of correlation but not to what extent one variable has a cauasl effect on another.

The inferential statistics do not support our initial hypothesis that there is a direct link between open and closed innovation. Still we can see from the analysis that both patents and GitHub users innovative activities require similar conditions, which indicates that there is an indirect link between the two variables of interest. These findings support the claims brought forward by Jaffe and Palmer (1997), Florida (2006) and Johnson (2014), who argue that cities need certain characteristics to attract creative talent and hence to become creative and innovative hubs. This is an important finding of our study, as it highlights the framework conditions for attracting and retaining talent at a certain locale, especially for such mobile talents working in programming and broader software developments (see Florida 2006). Still, when interpreting our findings,

one needs to aware that there might be other unaccounted for variables to contribute to patents and user counts but are not currently accounted for in the model.

Additionally, it seems relevant to reaffirm the limitations of the study, that even should this study have found a high correlation between GitHub and Patent data, or either of these variables and the explanatory variables, the dataset used only gives only a static relationship, rather than the relationship over time. Additionally it remains exceptionally difficult to ascertain a relevant and unarbitrary threshold at which to consider a user innovative. This study chose to make the threshold above 20 followers, largely because this stood as a clear break in the data, however, perhaps innovators are more or less common and more or less followed than we determine, and thus we exclude or include innovators unduly. Additionally, while we find no relationship between GitHub and patent data, perhaps this is due simply to the low diffusion of GitHub, which is still in its infancy, compared to the patent system, which is used the world over and has no competitors or alternative system.

5 Conclusion

The findings endorse the implicit hypothesis of the study that open data sources seem to show innovation in a similar but perhaps distinct manner to patent data and could hence enrich the measurement of innovation. While no direct correlation was found between patent data and GitHub user data, it remains an open question whether this means that follower data is a poor measurement for innovation, or whether potentially they measure a different type of innovation. It seems plausible that GitHub data offers a glimpse into the 'throughput' of open innovation (in the form of collaboration) rather than the 'output' which patent data reflects (in the form of commercialization of knowledge). It could be that patent data better takes into account the exact moment and type of innovation, as it is tied to the patent application itself, whereas someone could choose to follow someone based upon other reasoning than their level of innovation.

Interestingly, as the strongest relationships were found not between patent and follower data, but rather between these two variables and the city related control variables, that many of the conditions which foster both open and closed innovation, in this case GDP and city importance, both are strongly related to the

drive of innovation in a city. Due to these parallels, it seems plausible that policy makers could institute similar reforms to positively affect both. This research was done in response to claims made by scholars such as Taylor (2004) and Freeman and Soete (2009) who call for a rethinking of how we measure innovation.

Clearly our research approach relying solely on GitHub user data as a measure of open innovation is far from exhaustive. Further research could include other, similarly impotant collaborative research platforms. The fast diffusion of such research platforms could mean that similar analytical approaches could lead to much different findings. Other studies could also include more control variables to account for other factors influencing both patents and collaborative network activities, which due to limited data availability could not be analyzed in this research project. Furthermore, a time series or panel data analysis could look into differences between and within different cities over time and thereby significantly increase the validity of the research findings, also allowing for causal interpretations. However we decide to think of innovation, there is great need for new ways of measuring innovation to more accurately represent innovation in the open technological age.

References

Acs, Zoltan J, Luc Anselin, and Attila Varga. 2002. "Patents and Innovation Counts as Measures of Regional Production of New Knowledge." *Research Policy* 31 (7). Elsevier: 1069–85.

Bivand, Roger, and Nicholas Lewin-Koh. 2014. *Maptools: Tools for Reading and Handling Spatial Objects*. http://CRAN.R-project.org/package=maptools.

Blondel, Emmanuel. 2014. *Rsdmx: Tools for Reading SDMX Data and Metadata*. http://CRAN.R-project.org/package=rsdmx.

Breschi, Stefano, and Franco Malerba. 2005. *Clusters, Networks and Innovation*. Oxford University Press.

Brownrigg, Ray. 2014. *Maps: Draw Geographical Maps*. http://CRAN.R-project.org/package=maps.

Couture-Beil, Alex. 2014. *Rjson: JSON for R*. http://CRAN.R-project.org/package=rjson.

Eddelbuettel, Dirk, and Romain Francois. 2014. *Rcpp: Seamless R and C++ Integration*. http://CRAN.R-project.org/package=Rcpp.

Florida, Richard. 2006. "The Flight of the Creative Class." *Liberal Education*, 22–29.

Fox, John, and Sanford Weisberg. 2011. *An R Companion to Applied Regression*. Second. Thousand Oaks CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Frankelius, Per. 2009. "Questioning Two Myths in Innovation Literature." *The Journal of High Technology Management Research* 20 (1). Elsevier: 40–51.

Freeman, Christopher, and Luc Soete. 2009. "Developing Science, Technology and Innovation Indicators: What We Can Learn from the Past." *Research Policy* 38 (4). Elsevier: 583–89.

Gandrud, Christopher. 2014. *Repmis: A Collection of Miscellaneous Tools for Reproducible Research with R*. http://CRAN.R-project.org/package=repmis.

Gnyawali, Devi, and Manish Srivastava. 2013. "Complementary Effects of Clusters and Networks on Firm Innovation: A Conceptual Model." *Journal of Engineering Management*, no. 30: 1–20.

Hlavac, Marek. 2014. *Stargazer: LaTeX/HTML Code and ASCII Text for Well-Formatted Regression and Summary Statistics Tables*. http://CRAN.R-project.org/package=stargazer.

Jaffe, Adam B, and Karen Palmer. 1997. "Environmental Regulation and Innovation: A Panel Data Study." *Review of Economics and Statistics* 79 (4). MIT Press: 610–19.

Johnson, Mandy. 2014. *Winning the War for Talent: How to Attract and Keep the People to Make the Biggest Difference to Your Bottom Line*. John Wiley & Sons.

Kahle, David, and Hadley Wickham. 2013. *Ggmap: A Package for Spatial Visualization with Google Maps and OpenStreetMap*. http://CRAN.R-project.org/package=ggmap.

Pakes, Ariel, and Zvi Griliches. 1980. "Patents and R&D at the Firm Level: A First Report." *Economics Letters* 5 (4). Elsevier: 377–81.

Pianta, Mario. 2005. "Innovation and Employment." Oxford University Press.

Ponds, Roderik, Frank Van Oort, and Koen Frenken. 2010. "Innovation, Spillovers and University–industry Collaboration: An Extended Knowledge Production Function Approach." *Journal of Economic Geography* 10 (2). Oxford Univ Press: 231–55.

Porter, E., Michael. 1991. "America's Green Strategy." Scientific American.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Schumpeter, Joseph. 1942. "Creative Destruction." *Capitalism, Socialism and Democracy*.

Senghore, Fatima, Enrique Campos-Nanez, Pavel Fomin, and James S Wasek. 2014. "Using Social Network Analysis to Investigate the Potential of Innovation Networks: Lessons Learned from NASA's International Space Apps Challenge." *Procedia Computer Science* 28. Elsevier: 380–88.

Smith, K. H. 2005. "Measuring Innovation." PhD thesis, Oxford University Press.

Taylor, M. Z. 2004. "Empirical Evidence Against Varieties of Capitalism's Theory of Technological Innovation." *International Organization* 58 (03). Cambridge Univ Press: 601–31.

Temple Lang, Duncan. 2014. *RCurl: General Network (HTTP/FTP/.) Client Interface for R*. http://CRAN.R-project.org/package=RCurl.

Vaidyanathan, Ramnath. 2013. *RCharts: Interactive Charts Using Javascript Visualization Libraries*.

Wickham, Hadley. 2014a. *Httr: Tools for Working with URLs and HTTP*. http://CRAN.R-project.org/package=httr.

———. 2014b. *Reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package*. http://CRAN.R-project.org/package=reshape2.

Wickham, Hadley, and Winston Chang. 2014. *Ggplot2: An Implementation of the Grammar of Graphics*. http://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, and Romain Francois. 2014. *Dplyr: A Grammar of Data Manipulation*. http://CRAN.R-project.org/package=dplyr.