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

Benjamin Snow and Oliver Bott

14 November 2014

1 Background

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?*

For the full examination of the previous literature on the subject and reasoning for the purpose, motives, and plan for this study, please see the Research Proposal. This contribution focuses specifically on outlining the data gathering, data cleaning, and merging process. It will also examine the constructed dataset using basic descriptive statistics, as well as run preliminary inferential statistical models, offering explanation and context throughout the process. Finally, steps to improve the analysis for the final version will be given.

2 Data Gathering

To examine open network data against patent data, this study relies on two data key sources and uses the statistical tool *R* (R Core Team 2014) for the data analysis.

The first data set was obtained by using the Application Programming Interface (API) data for open networks. To examine open data innovation, data is obtained from the the git repository web-based hosting service GitHub[1]. The *R* (R Core Team 2014) packages *httr* Wickham (2014), *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 reponsitories.

As an indicator of closed innovation we use city-level patent registration data from the Organization for Economic Co-operation and Development[2]. We use PCT patent applications per 10,000 inhabitants. From the same database, we also use GDP per capita data and environmental data, as other variables which could prove significant in explaining differences in innovation. From the same OECD database: GDP per capita, pollution levels, greenspace, and employment (as share of national employment) are taken for the relevant cities which github and patent data are obtained for. These other variables are inclued in the analysis as they could prove significant in explaining differences in innovation. The *R* (R Core Team 2014) package *rsdmx* (Blondel 2014) is necessary for obtaining the OECD dataset.

All data obtained via the GitHub API and OECD database can be linked to individual cities, allowing for an analysis on the regional level. The code used for gathering and cleaning the data can be accessed via this link.

3 Data Sources

Table 1 depicts the variables used in the study.

[1] Online accessilbe via https://github.com/.

[2] Online accessible via http://stats.oecd.org.

Variables	Year	Source
Patents	2008	OECD
GDP	2008	OECD
Population	2008	OECD
Greenspace	2008	OECD
Employment	2008	OECD
Pollution	2005	OECD
No Following	2014	GitHub API
1-24 Following	2014	GitHub API
>25 Following	2014	GitHub API

4 Prelimianary Descriptive Statistical Analysis

A first examination of the distribution of each variable show distributions which reflect a . . . .. An examination of the distribution of both the patent applications and github users in each city are taken to examine if a similar distribution can be seen in cities when normalizing for population size. A similar distribution would indicate . . . (Olli, question here, if its examining when it is only no following, this isnt giving us total user numbers, just user numbers of those who dont have any following, right?)

Table 2: Summary statistics

	X	METRO_ID	Patents	GDP	Population	Greenspace	Employment	Pollution	nofollowin
4	4	Albuquerque	0.73	38,060.00	852,670	71.24	0.27	8.55	3.60
5	5	Amsterdam	1.18	40,064.00	2,317,340	219.30	14.13	22.57	8.23
6	6	Antwerp	1.08	38,605.00	1,040,008	318.70	10.29	23.02	1.89
7	7	Atlanta	1.27	42,352.00	4,202,262	111.10	1.42	16.20	6.53
8	8	Augsburg	2.10	33,636.00	594,054	752.00	0.78	20.83	1.40
9	9	Austin	2.68	48,859.00	1,608,564	172.00	0.51	9.67	23.16
10	10	Baltimore	1.32	61,804.00	1,939,607	1,000.00	0.68	20.65	3.91
11	11	Barcelona	0.93	33,874.00	3,594,393	2.45	8.69	14.88	5.64

Next, the *car* package Fox and Weisberg (2014) 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. Taking into account the clumped distribution of many variables, taking the log regression of several variables when regressing is attempted, to see if this brings about a more normal distribution.

5 Inferential Statistics

This study plans to use an ordinary least squares (OLS) model to examine the relationship between patent and highly followed open data sources. The high level of significance between these two variables, rather than demonstrating that one is acting on the other, show that some other 'spur of innovation' is acting on both, but does endorse the implicit hypothesis of the study that open data sources do seem to show innovation in a similar but perhaps distinct manner to patent data. This found relationships offers a glimps into the 'throughput' of innovation rather than the 'output' which patent data reflects. However, this is exceptionally early analysis and will need to be examined further.

Now that a relationship between patent and follower data has been shown, further inferential statistical analysis attempts to find the common predictor or cause of innovation in both patent and open data is necessary. This is done by running identical Ordinary Least Squares regressions, using the the log of both

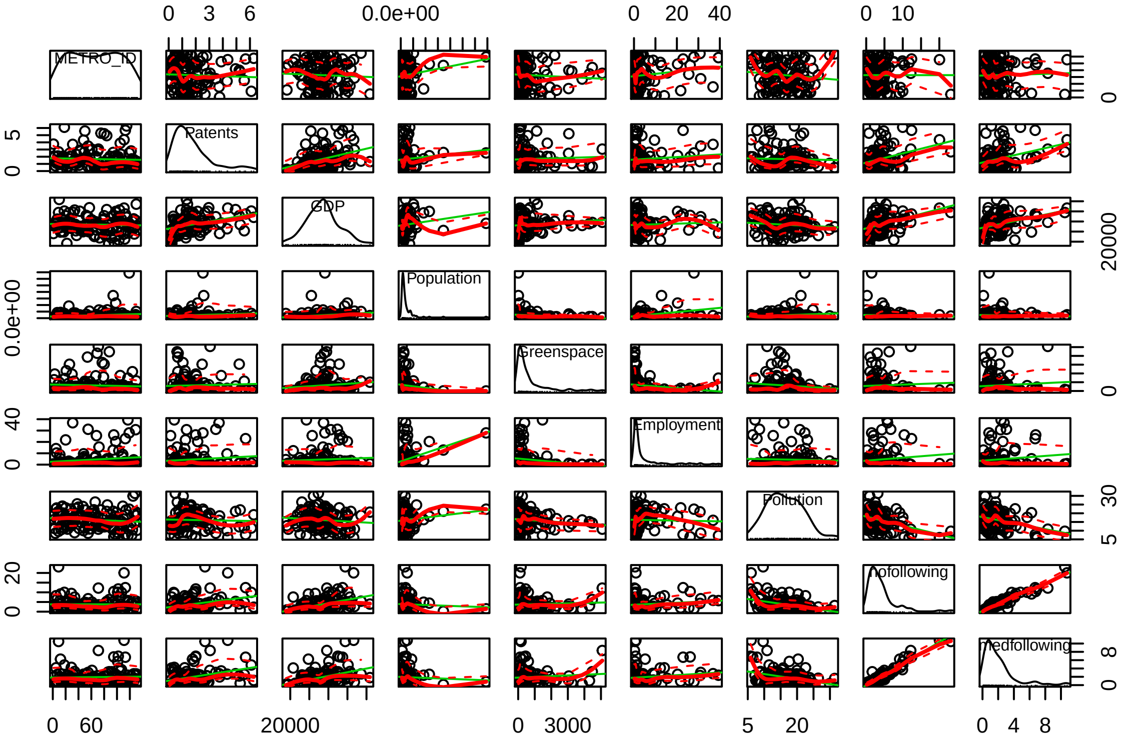

Figure 1: plot of chunk unnamed-chunk-5

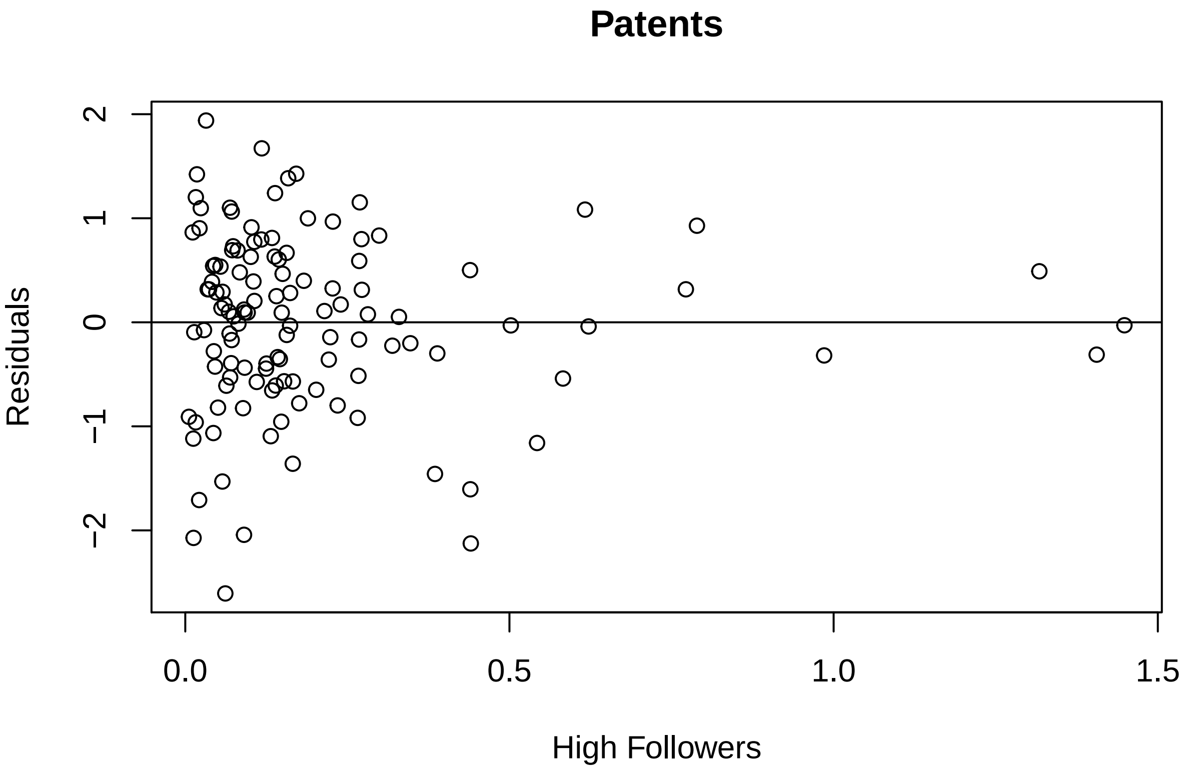

Figure 2: plot of chunk unnamed-chunk-6

patent and highly followed github users as the dependant variables, and the log of each of the previously mentioned independant variables (GDP, Population, Greenspace, Employment, and Pollution), to examine if any of these variables prove significantly (and presumably, similarly) correlated to our dependant variables.

This process has allowed us first directly compare our previously established (patent) and newly hypothesized (open - github) measures of innovation. After comparing directly, attempting to find how they differently reflect innovation, as well as attempting to find the common cause of innovation for both the innovation throughput which is open data, and the innovation output which is patent data, is necessary.

Table 3: Regression Estimates of Patent Activity

	Dependent variable:			
	log(Patents)			
	(1)	(2)	(3)	(4)
(Intercept)	0.32*** (0.08)			
No Following		0.34*** (0.08)		
1-24 Following			0.34*** (0.07)	0.15* (0.08)
>25 Following				1.76*** (0.41)
GDP				−0.06 (0.11)
Population				0.05 (0.05)
Greenspace				0.08 (0.06)
Employment				0.49** (0.21)
Pollution	−0.13 (0.11)	0.10 (0.08)	0.89*** (0.17)	−18.73*** (3.92)
Observations	120	120	120	120
R^2	0.11	0.15	0.15	0.34
Adjusted R^2	0.10	0.14	0.15	0.31
Residual Std. Error	0.87 (df = 118)	0.85 (df = 118)	0.85 (df = 118)	0.76 (df = 113)
F Statistic	14.72*** (df = 1; 118)	20.50*** (df = 1; 118)	21.58*** (df = 1; 118)	9.88*** (df = 6; 113)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4: Regression Estimates of Follower Numbers

	Dependent variable:	
	log(hifollowing)	
	(1)	(2)
(Intercept)	0.46*** (0.10)	0.18* (0.10)
Patents		2.21*** (0.45)
GDP		−0.19 (0.12)
Population		0.04 (0.05)
Greenspace		0.09 (0.07)
Employment		−0.65*** (0.23)
Pollution	−2.21*** (0.09)	−21.19*** (4.38)
Observations	120	120
R^2	0.15	0.39
Adjusted R^2	0.15	0.36
Residual Std. Error	0.98 (df = 118)	0.85 (df = 113)
F Statistic	21.58*** (df = 1; 118)	12.01*** (df = 6; 113)
Note:		*p<0.1; **p<0.05; ***p<0.01

References

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

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

Fox, John, and Sanford Weisberg. 2014. *Car: Companion to Applied Regression*. http://CRAN.R-project.org/package=car.

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

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

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