

# *Coding Together - Coding Alone:* The Role of Trust in Collaborative Programming

Fabian Stephany<sup>a</sup>, Fabian Braesemann<sup>b\*</sup>, Mark Graham<sup>b</sup>

<sup>a</sup>Vienna University of Economics and Business

<sup>b</sup>Oxford Internet Institute - University of Oxford

## ARTICLE HISTORY

Compiled August 24, 2019

## ABSTRACT

In the digital economy, innovation processes increasingly rely on highly specialised know-how and open-source software shared on digital platforms on collaborative programming. The information that feeds into the content on these platforms is provided voluntarily by a vast crowd of knowledgeable users from all over the world. In contributing to the platforms, users invest their time and share knowledge with strangers to add to the rising body of digital knowledge. This requires an open mindset and trust. In this study, we argue that such a mindset is not just an individual asset, but determined by the local communities the users are embedded in. We, therefore, hypothesise that places with higher levels of trust should contribute more to Stack Overflow, the world's largest question-and-answer platform for programming questions. In relating the city-level contributions of 266 OECD metropolitan areas to infrastructure, economic, and trust measures, we find this hypothesis confirmed. In contrast, click rates to the platform are solely driven by infrastructure and economic variables, but not by trust. These findings highlight the importance of societal values in the 21st century knowledge economy: if policy-makers want to develop a lively local digital economy, it is not enough to provide fast Internet access and business opportunities. Instead, it is equally important to establish a trust-building environment that fosters sharing of innovative ideas, collaborations, and knowledge spillovers.

## JEL CLASSIFICATION

J 24, L 17, O 35, R 12

## KEYWORDS

Digital Knowledge Economy • Internet Geography • Platform Innovation • Collaborative Programming • Stack Overflow • Trust

## 1. Introduction

In the 21st century, digital platforms have become ubiquitous tools for the transfer of information and the distribution of knowledge. Every day, millions of people consult *Wikipedia*, the world's largest encyclopedia; numerous users download open-source software from the project hosting website *GitHub*, and they browse the question-and-answer website *Stack Overflow* to get answers to all kinds of programming questions.

The vast amount of information that feeds into the articles, software products, and threads on these platforms is collaboratively produced by a large crowd of knowledgeable users from all over the world. In contributing to the platforms, the users add to the rising body of digital knowledge. Consequently, this form of platform-enabled 'distributed innovation' (Yoo et al., 2012) is becoming an increasingly important element of innovation processes.<sup>1</sup>

---

\* Email: fabian.braesemann@sbs.ox.ac.uk, ORCID-ID: 0000-0002-7671-1920

<sup>1</sup>In this study, we do not claim that platform contributions *are* innovations, in the sense that these activities were focused

As these platform-based innovative activities happen in the digital sphere, they are in principle not bound to local knowledge clusters that characterised earlier forms of innovation. Traditionally, innovation has been concentrated in cities and metropolitan areas, where the physical proximity of people enables a vivid informal exchange of ideas (Carlino et al., 2007; Marshall, 2009), facilitated by universities and other research institutions (Andersson et al., 2009; Lin, 2011; Forman and Zeebroeck, 2012).

While the reduced search and communication costs associated with digital technologies (Forman et al., 2018) have certainly simplified the coordination of distributed innovation, the 'Information Superhighway' (Sawhney, 1996) has not turned the world into a 'global village' (McLuhan, 1964). Digital knowledge production and collaborative programming require specialised know-how, and the users that are able to contribute to the platforms are embedded in their local communities. As a consequence, the geography of platform contributions is likely to follow the uneven distribution of economic resources, infrastructure, and human capital (Braesemann et al., 2019; Stephany and Braesemann, 2017; Graham et al., 2015).

Moreover, platform activities like collaborative programming lack formal enforcement mechanisms such as contracts or monetary rewards: the users share their knowledge voluntarily and they invest their time to help others (Penoyer et al., 2018). We, therefore, argue that the 'social embeddedness' (Granovetter, 1985) of users in trust-building local communities plays an important role in understanding the geography of collaborative programming contributions.

While the role of trust in online communities has been analysed on the individual or group-level (Kobayashi et al., 2006; Diamant-Cohen and Golan, 2017), it has not been investigated whether regional values of societal trust are related to the large-scale spatial patterns of digital platform contributions.

In this study, we aim to close this gap in investigating data from Stack Overflow, the world's largest questions-and-answer platform for programming. We hypothesise that places with higher levels of societal trust<sup>2</sup> will more intensively contribute to the platform, as voluntary contributions do not only require Internet access and human capital, but also the willingness to voluntarily engage with others. In contrast, the consumption of digital knowledge does not require specific societal values. We therefore hypothesise that the regional levels of Stack Overflow visits should not be affected by societal trust.

To test the hypotheses, we assign more than two million users and 100 million visits to countries and metropolitan areas, and we compare the global distribution of website traffic and contributions on the country-level, and on the city-level for 266 OECD metropolitan areas. The results suggest that both the highly unevenly distributed clicks and contributions to Stack Overflow can be associated to economic and digital infrastructure imbalances, while, in line with our theoretical framework, only contributions tend to be related to societal trust.

In the following, we present an overview of the literature on the economic geography of innovation, on platform innovation, and on the relation to trust in section 2, from which we derive the research hypotheses in section 3. The research design is outlined in section 4, the results are described in section 5, and the last section concludes.

---

on marketable outcomes, but they are a form of digital knowledge creation. We relate platform contributions to the patent-based innovation literature, as it provides a guideline on how the geography of such activities can empirically be modelled.

<sup>2</sup>At the same time, a society's level of generalized trust is strongly related to societal settings, such as inequality (Stephany, 2017) or educational attainment (Stephany, 2019).

## 2. Related Work

### *Patent-based Innovation and the City*

Researchers from economics, management and organisational studies have examined the geography of innovation from various perspectives (Carlino and Kerr, 2015). There is consensus that innovation, as measured by patent activity, most often takes place in densely populated regions, like cities or metropolitan areas (Carlino and Kerr, 2015; Carlino et al., 2007; Florida and Mellander, 2018; Forman et al., 2016). Usually, four concepts are used to explain the concentration of innovation in cities:

First, the high density of population and employment allows to match highly skilled, specialised labour demand and supply more easily (Fallick et al., 2006; Carlino et al., 2007).

Secondly, knowledge spillovers — the informal exchange of ideas due to the physical proximity of people — are a highly localised phenomenon (Carlino and Kerr, 2015). Innovative corporations like Bell Labs or Google have explicitly designed their headquarters to increase the physical interactions of people in order to use these local spillovers and to increase the exchange of ideas (Gertner, 2012).

Thirdly, urban areas have an institutional advantage: universities and other research facilities, which are often located in cities, explicitly encourage inventive activities and knowledge spillovers (Carlino et al., 2007; Forman et al., 2014). For instance, Andersson et al. (2009) show how the reallocation of universities can revitalise regional innovation, and Lin (2011) find that the degree of urbanisation and university locations matter for local innovation.

Fourthly, cities have been found to attract and empower talent because of higher levels of cultural openness, migration, and societal trust: Ozgen et al. (2012) report that the diversity of immigrant communities is beneficial for inventive output in European regions. Likewise, Nathan (2015) show that the diversity of inventor communities helps to raise individual patenting activities in the UK. Several studies find that innovation benefits from high levels of trust, strong cooperative norms, and extended networks (Akomak and Weel, 2009; Doh and Acs, 2010; Westlund and Adam, 2010).

### *Platform Innovation and Digital Knowledge Creation*

Today, digital technologies and, in particular, online platforms fundamentally affect how innovation processes are organised (Yoo et al., 2012). Companies that provide data and services via platforms gain additional business opportunities, as other firms integrate these platforms in creating new products (Hauptman, 2003; Gawer, 2009). Platforms disperse the innovation process to the periphery of organisations and towards open-source communities (Boudreau, 2010), and they foster combinatorial innovation (Faraj et al., 2011).

While digital innovation processes differ from more traditional forms of innovation, studies have long been solely focused on patent activity as a measure of innovation. Only recently have scholars started to investigate platform-based knowledge creation and innovative activities, or the emerging economic geographies associated with these phenomena.

Graham et al. (2015) map the global editing activities of Wikipedia and find that the creation of digital knowledge on the platform is determined by economic factors and Internet infrastructure. With a focus on US metropolitan areas, Stephany and Braesemann (2017) show that Wikipedia edit activities on topics around information and communication technologies are

concentrated in cities with an academic computer science department.

Likewise, open source software development contributions are geographically clustered: most of the users of the popular project hosting website GitHub are located on the US coastlines or in Europe (Lima et al., 2014; Takhteyev and Hiltz, 2014), and the majority of user connections take place within a 300 km radius. As with Wikipedia contributions, the global distribution of online collaborations in software development is correlated with Internet infrastructure and economic factors (Gonzalez-Barahona et al., 2008; Engelhardt et al., 2013). Davidson and Poor (2018) examine the role of crowd-funding platforms in the dispersion of innovative projects in the United States and find that online crowd-funding projects tend to be located in cultural hubs.

The research approaches of several empirical investigations on the geography of patent-based innovation and digital knowledge creation are summarised in Table 1. While measures on Internet infrastructure, GDP, demographics, and the stock of human capital have been widely employed in both strands of research, societal values or trust have so far only been considered in empirical studies on the geography of patent activity, but not to explain platform contributions.

Table 1 near here.

### *Digital Platform Contributions and Trust*

One important feature of platform-mediated collaborative activities is the lack of formal enforcement mechanisms (Boudreau, 2010): Collaborative digital knowledge creation is not governed by contracts and it does not yield monetary rewards. Instead of using these market-based incentives, the platforms establish cooperative norms of trustworthy behaviour and reputation systems as enforcement mechanisms in order to 'socially embed' (Granovetter, 1985) the user interactions.<sup>3</sup>

English-Lueck et al. (2002) point out that 'using technologically-mediated communication, requires a high degree of trust' (p.90). In their revision of anthropological studies of the last 20 years on Silicon Valley tech-communities, they underline trust's important mediating role: 'high-tech knowledge work is done by networks of interdependent global workers that must share information, act under a severe time constraint, and establish effective relationships at a distance'.

Similarly, Diamant-Cohen and Golan (2017) investigate on the role of trust in platform-based file-sharing communities. They describe active platform contributors as members of a 'goal-oriented community [...] in opposition to dominant corporate platforms' (p.1737), and point out that the success of peer-collaboration in the analysed communities is strongly driven by trusting 'physically anonymous counterparts' and solidarity.

O'Neil (2014) emphasise the fragility of trust in digital peer production in light of 'enhanced verifiability'. Actors in such environments might be at risk of losing the trust of the community, since the results of their work can be easily verified by the crowd ('either the computer code runs, or it doesn't', p. 882).

Kobayashi et al. (2006) show that the reciprocity of generalised trust (trust emerges from the experience of trustful collaborations) can be extended to digital interactions. Their study of 1,300 adults in Japan exhibits that collaboration and interactions via the Internet simultaneously require and enhance 'online trust'. They find spillovers from trustful online interactions

---

<sup>3</sup>For example, Wikipedia editors can gain awards for editing; on GitHub, a user's contribution history is publicly displayed on the profile page; and on Stack Overflow, users 'earn' reputation points for contributions (Bosu et al., 2013).

to generalised trust in the 'offline' world: Individuals with high levels of 'online trust' are more engaged in their civic society.

### 3. Hypotheses

Based on the literature reviewed in the previous section, we argue that the geography of interactions with platforms of the digital knowledge economy, in this study measured by the regional number of clicks and contributions to Stack Overflow, should be influenced by infrastructure and economic factors. To contribute to the platform, one needs specific skills that allow to pose unanswered questions or to reply to such questions. Users with such highly specialised knowledge should be more likely located in places with a strong local economy and internet infrastructure. This is captured by the first hypothesis:

*H1: On the country- and city-level, contributions and clicks to Stack Overflow are positively associated with internet infrastructure and economic factors.*

Additionally, we argue that the geography of digital platform *contributions* do not only mirror regional differences in economic specialisation and infrastructure, but also regional differences in the societal attitudes towards trust, because people need to trust others to share knowledge and to interact voluntarily with strangers on digital platforms. Such attitudes are commonly measured by the *generalised trust* question.<sup>4</sup> In contrast to clicks, regional levels of contributions should thus be affected by societal trust:

*H2: Places with higher levels of generalised trust contribute more to Stack Overflow than places with lower levels of trust.*

*H3: This relation does not hold for clicks to the website.*

### 4. Materials and Methods

#### *Data Sources*

To investigate the research hypotheses, we use a dataset of more than 19 million contributions to Stack Overflow from more than two million users.<sup>5</sup> Additionally, we collected Stack Overflow click data from *Quantcast*, a platform that tracks website traffic.<sup>6</sup>

We have restricted the dataset to user contributions made in 2014 and 2015, in accordance with the most recent completely available data sets from the World Bank (country-level controls)<sup>7</sup> and OECD (city-level controls).<sup>8</sup> Information on generalised trust is retrieved from the 2014–2016 waves of the World Value Survey.<sup>9</sup>

---

<sup>4</sup>The level of generalised trust is measured by the share of individuals, who agree to the question 'Generally speaking, do you think that most people can be trusted (or do you think that you can never be too careful in dealing with people)?'

<sup>5</sup>All Stack Overflow data are publicly available: <https://archive.org/details/stackexchange>. Details are described in (Braesemann et al., 2019).

<sup>6</sup><https://www.quantcast.com/stackoverflow.com>.

<sup>7</sup>World Bank: <https://data.worldbank.org/>

<sup>8</sup>OECD metropolitan area dataset: <https://stats.oecd.org/Index.aspx?Datasetcode=CITIES>

<sup>9</sup><http://www.worldvaluessurvey.org/wvs.jsp>.

## Data Analysis

We provide descriptive statistics and apply multivariate regression models to test the research hypotheses. While descriptive statistics are provided on both the national- and city-level, we focus the inferential analysis on cities, acknowledging the importance of local geographies in understanding innovative activities. Due to a lack of availability and comparability of demographic and economic city-level measures on a global scale, we limit this part of the analysis to 266 metropolitan areas in OECD countries. Following the research approaches of the empirical studies on traditional and digital innovation geographies summarised in Table 1, we relate the city-level counts of Stack Overflow contributions and clicks to population size and density, GDP per capita, employment rate, broadband access, affinity<sup>10</sup>, population share with at least secondary education, age-dependency-ratio, and to the level of trust. Definitions and descriptive statistics on all these variables are provided in Table 2 in the Appendix.

The choice of control variables in the final regression model is the result of a step-wise forward and backward model selection procedure (see Fig. 3A in the Appendix). To validate the model, different specifications have been compared and the robustness of the final model has been assessed by its' out-of-sample cross-validated prediction accuracy (Fig. 3B and 3C).

## 5. Results

### *Global Distribution of Stack Overflow Contributions and Clicks*

Contributions and clicks to Stack Overflow are geographically clustered between and within countries. Figure 1 summarises the global geographies: panel A shows the number of Stack Overflow contributions in 188 countries (the most active city per country is highlighted by circles of different sizes according to the per-capita contributions) and the national levels of generalised trust (country colour). Cities with higher values of per-capita contributions are located in Europe, North America and India, where trust on average is high. Less contributing cities tend to be located in countries with lower levels of trust (predominantly in Africa, South America, and East-Central Asia).

Figure 1 near here.

Contributions are very unevenly distributed between countries and cities (Fig. 1B): while the majority of cities in the dataset have less than 250 contributions, the most active 20 % of the cities account for nearly 85 % of all metropolitan-level contributions. Users from the three most active cities, New York, London, and Bangalore, together (1.5 % of the cities) account for 131,000 contributions - more than all contributions of the 90 % least active cities taken together. Such a fat-tailed distribution is typical for activity patterns of online platforms (Graham et al., 2015; Davidson and Poor, 2018). Clicks to Stack Overflow show a similarly skewed distribution

---

<sup>10</sup>Affinity describes how likely an Internet user in a given city is to visit Stack Overflow, relative to the global average of all Internet users. For example, in a world with only two cities, *A* (1,000,000 Internet users and 1,000 Stack Overflow visitors) and *B* (5,000,000 Internet users and 2,000 Stack Overflow visitors), city *A* would have an affinity value of 2.0 and city *B* a value of 0.8. Thereby, affinity does not depend on the size of a city's Internet population: it captures differences in Internet infrastructure and general affinity of the population to use the Internet; in other words the variable is used in this study as a normalised measure of the topical affinity to or *specialisation* in programming content on the Internet. This variable should, on its own, capture a lot of the variation of Stack Overflow contributions and clicks and helps us to control for some of the unobserved heterogeneity that is due to differences in the use of the Internet.

Source: <https://www.quantcast.com/measure/stackoverflow.com?country=US#/generalInterestsCard>

and closely coincide with contributions on both the city- and country level (Fig. 1C). However, contributions and clicks are not identical.

Previous studies on digital innovation show that Internet access and GDP are important factors for local platform activity (Graham et al., 2015; Davidson and Poor, 2018; Engelhardt et al., 2013). We therefore compare the country contributions to Stack Overflow with affinity in Figure 2A. The upper panel shows the Stack Overflow contributions per country grouped in the four quartiles of each of the three variables. A clear upwards trend is visible in all three cases, while difference between bottom and top groups are most significant for GDP: countries in the highest GDP quartile count on average almost 100 times more contributions than counterparts in the lowest GDP quartile.

Figure 2 near here.

### *OECD Cities and Robustness Checks*

The observations made so far provide descriptive evidence in favour of the first hypothesis. However, for more detailed comparisons between socio-economic regional factors and Stack Overflow contributions, we focus on the subset of 266 OECD cities. In the lower panel of Figure 2A, the OECD cities are, again, grouped by levels of affinity, GDP, and generalised trust. Here, in addition to affinity and GDP, quartile differences for trust are more pronounced: OECD cities in the highest trust group have on average ten times as many contributions as metropolitan areas in the bottom trust-quartile.

The extreme right-skewed distribution of contributions is our starting point for the selection of a suitable regression model. An additional issue for the selection of an adequate model is that for 22 cities no contributions had been detected. We assign zero values to these cities. In order to normalise the outcome distribution, we apply an inverse hyperbolic sine transformation.<sup>11</sup>

In a next step, we select relevant control variables from a comprehensive list of features (Tab. 2) in a standardised way, by performing forward and backward step-wise model selection based on the adjusted  $R^2$  of the model (Fig. 3A). Both forward and backward step-wise model selection yield one optimal model describing the OECD city-level contributions to Stack Overflow. The chosen variables contain the logarithmic values of affinity, population size, population density, GDP per capita, employment rate, and trust.

In order to validate the model specification, we compare the chosen inverse hyperbolic sine transformed OLS model with alternative model specifications (see Fig. 3B and 3C). There are no substantial differences with regards to the inclusion or exclusion of the zero-count cities (comparison between models 1. and 2.; and between 3. and 4.). We therefore stick to our choice of assigning zero values to cities from which no contributions were detected. Moreover, we conclude that the negative binomial models (3. and 4.) yield very similar results compared to the linear models. However, the negative binomial models are less accurate in terms of cross-validated out-of-sample prediction accuracy (Fig. 3C). By applying leave-one-out cross validation, we control for the effects of outliers and detect that the negative binomial models fail to accurately predict extreme contribution values (Fig. 3D). We therefore conclude that the inverse hyperbolic sine transformed model (1) is best suited to describe the number of city-level Stack Overflow

---

<sup>11</sup> $y = \log(x + \sqrt{x^2 + 1})$ . This transformation is an alternative to the more commonly used logarithmic transformation to handle extreme values, but has the advantage of being defined for zero-values, see (Burbidge et al., 1988).

contributions in terms of socio-economic local factors.

### *Regression Results*

In light of the model selection procedure and robustness checks, we explain the normalised values of Stack Overflow contributions in a linear model environment (Fig. 2B and 2C). Model (1) shows that affinity has a strong positive association with the number of contributions. Not surprisingly, the more likely a city’s individuals are to visit pages with programming content, the higher their likelihood to contribute to Stack Overflow. An additional, intuitive result is that population size affects contributions positively: the larger a city’s population, the higher the number of contributions. The negative coefficients for population density and employment rate appear counter-intuitive at first sight. However, both coefficients are likely to be driven by two ‘outlier’ countries: Japan and Korea are both densely populated and enjoy high employment, but they show relatively low levels of Stack Overflow contributions. Economic output, on the other hand, is positively related to contributions. Lastly, the level of generalised trust shows a positive association with city-level contributions. Besides aspects of infrastructure, population or economic characteristics, cities with higher levels of trust, tend to contribute more to Stack Overflow. These findings confirm our first two research hypotheses.

Figure 2B moreover relates clicks to Stack Overflow to city-level characteristics in model (2), which has also been optimised by step-wise model selection. While the model yields similar results, trust is not selected as a substantially explanatory variable. In model (3), we have included trust as additional factor, but neither is the variable statistically significant, nor does it improve explanatory power in terms of  $R^2$ . This finding is in line with our third hypothesis: As clicks to Stack Overflow represent only an individual act of digital knowledge consumption, they are not affected by cooperative norms, such as generalised trust.<sup>12</sup>

## **6. Conclusion**

Online platforms have given rise to a new form of decentralised, re-combinatorial, and collaborative form of innovation. Thereby, many platforms rely on an open-source creation process with no formal enforcement of collaboration, such as monetary incentives or contracts. Instead, cooperative norms, like reputation or trust, support the networks of open-source communities. Our work examines the Q&A coding forum Stack Overflow as an example of global open-source collaborative programming. User contributions to the platform are identified as items of digital knowledge production, while clicks to Stack Overflow represent digital knowledge consumption. We have found that trust plays an important role in explaining the geography of contributions, but not in explaining click rates.

Digital development policies, like the UN’s ‘ICT for SDG<sup>13</sup> initiative’<sup>14</sup>, underline the importance of Internet infrastructure investments for developing countries to join the global digital

---

<sup>12</sup>The conceptualisation and cross-cultural comparability of generalised trust has been criticised in the past by Sobel (2002) and others (Stephany et al., 2017). For the case of collaborative programming activities, we were interested in the willingness to trustworthy collaborate with strangers; a concept that we have approximated by generalised trust. In order to validate this presumption, we run a separate analysis in which we use the share of a city’s population engaged in charitable organisations as a control instead of trust. Similarly to the results reported here, we observe a strong relationship of charitable engagement and contributions, while clicks and voluntary engagement are only weakly related.

<sup>13</sup>Sustainable Development Goals

<sup>14</sup><https://www.itu.int/en/sustainable-world/Pages/default.aspx>



knowledge economy. Certainly, broadband connectivity and local digital business opportunities need to be improved in order for currently disadvantaged regions to catch up (Tjoa and Tjoa, 2016). However, our research findings indicate that, besides economic performance and infrastructure, societal trust is an essential ingredient in the development of an active local hub of people willing to contribute to knowledge-building collaborative programming platforms.

Thus, for the establishment of a lively programmer community and in order to make regions better equipped to participate in the global digital knowledge economy, policy-makers should not only concentrate on the provision of fast Internet, but also aim to establish a trust-building, innovation-friendly environment; one that fosters sharing of knowledge and interactions, and not one that isolates.

### *Limitations*

While we could not control for differences in the political environment or Internet censorship, we speculate that such regimes will rather tend to harm the voluntary engagement of innovative programmers on online platforms than to foster them. As the positive effects of these contributions in knowledge building are known (Moy et al., 2010; Vasilescu et al., 2013), and as building reputation on such systems is important for a programmer’s future career (Capiluppi et al., 2013; Penoyer et al., 2018), societies that harm trust-building might find it difficult to establish innovative digital economies that are built on an open-source sharing culture.

Stack Overflow is certainly the world’s most referenced Q&A platform for programming, but it is not the only forum on which programmers can exchange with each other. Some regional communities, might opt for other, smaller and potentially local, platforms to communicate between peers. In our work, we do not claim to holistically capture all interactions between programmers and developers from all regions of the world, but rather label Stack Overflow as the single largest website on collaborative programming on a global scale.<sup>15</sup>

---

<sup>15</sup>It could be, for instance, that users from countries where English is not so commonly used, rather prefer to use programming platforms in other languages. However, an exploratory search on [www.similarweb.com](http://www.similarweb.com), a website to measure country-specific web traffic, conducted on in March 2019, shows that Stack Overflow is the number-one platform in the category ‘Computer and Electronics’ in 54 out of 57 countries listed on SimilarWeb. In Japan it is ranked fifth, in Russia and Colombia ranked second. In all OECD, but Japan, it is ranked first. Considering the high correlation ( $\beta = 0.81$ , see Fig. 1C) between Stack Overflow clicks and contributions, and the high popularity of the website globally and across OECD countries, we conclude that language barriers (even if we do not control explicitly for such barriers in the analysis) should, thus, be of minor importance in explaining the global distribution of Stack Overflow contributions.

## **Acknowledgements**

This work benefited from comments made at the Complexity Science Hub Vienna Winter School 2019 in Obergurgl, and remarks from participants of the Innovation and Entrepreneurship Group Meeting at the Alexander von Humboldt Institute for Internet and Society in March 2019 in Berlin.

## **Disclosure statement**

Not applicable.

## **Funding**

This research was in part supported by the European Research Council grant number 335716.

## **Notes on contributors**

Fabian Stephany is a free-lance Computational Social Scientist, a Researcher at the Wittgenstein Centre for Demography and Global Human Capital, Vienna, and an Affiliate at the Humboldt Institute for Internet and Society in Berlin. His research focuses on the application of data science and social science statistics in education, migration, and public policy.

Fabian Braesemann is Research Fellow and Data Scientist at the Saïd Business School and Research Associate at the Oxford Internet Institute, University of Oxford. His research focuses on data mining and the statistical analysis of large-scale online data to understand market and information dynamics in a digitally connected world.

Mark Graham is Professor of Internet Geography at the Oxford Internet Institute, a Faculty Fellow at the Alan Turing Institute, a Senior Research Fellow at Green Templeton College, and an Associate in the University of Oxford's School of Geography and the Environment. He leads a range of research projects spanning topics between digital labour, the gig economy, internet geographies, and ICTs and development; with Digital Geographies being his most long-standing research area.

## References

- Akomak, S., Weel, B. t., Jul. 2009. Social Capital, Innovation and Growth: Evidence from Europe. *European Economic Review* 53 (5), 544–567.
- Andersson, R., Quigley, J. M., Wilhelmsson, M., Jul. 2009. Urbanization, Productivity, and Innovation: Evidence from Investment in Higher Education. *Journal of Urban Economics* 66 (1), 2–15.
- Bosu, A., Corley, C. S., Heaton, D., Chatterji, D., Carver, J. C., Kraft, N. A., 2013. Building Reputation in Stack Overflow: An Empirical Investigation. In: 2013 10th Working Conference on Mining Software Repositories (MSR). IEEE, pp. 89–92.
- Boudreau, K., 2010. Open Platform Strategies and Innovation: Granting Access vs. Devolving Control. *Management Science* 56 (10), 1849–1872.
- Braesemann, F., Stoehr, N., Graham, M., 2019. Global Networks in Collaborative Programming. *Regional Studies, Regional Science* (Forthcoming).
- Burbidge, J. B., Magee, L., Robb, A. L., 1988. Alternative Transformations to Handle Extreme Values of the Dependent Variable. *Journal of the American Statistical Association* 83 (401), 123–127.
- Capiluppi, A., Serebrenik, A., Singer, L., 2013. Assessing Technical Candidates on the Social Web. *IEEE software* 30 (1), 45–51.
- Carlino, G., Kerr, W. R., Jan. 2015. Agglomeration and Innovation. In: Duranton, G., Henderson, J. V., Strange, W. C. (Eds.), *Handbook of Regional and Urban Economics*. Vol. 5 of *Handbook of Regional and Urban Economics*. Elsevier, pp. 349–404.
- Carlino, G. A., Chatterjee, S., Hunt, R. M., May 2007. Urban Density and the Rate of Invention. *Journal of Urban Economics* 61 (3), 389–419.
- Davidson, R., Poor, N., Mar. 2018. Location, Location, Location: How Digital Platforms Reinforce the Importance of Spatial Proximity. *Information, Communication & Society* 25 (1), 1–15.
- Diamant-Cohen, A., Golan, O., Nov. 2017. Downloading Culture: Community Building in a Decentralized File-Sharing Collective. *Information, Communication & Society* 20 (11), 1737–1755.
- Doh, S., Acs, Z. J., Jun. 2010. Innovation and Social Capital: A Cross-Country Investigation. *Industry & Innovation* 17 (3), 241–262.
- Engelhardt, S. v., Freytag, A., Schulz, C., Apr. 2013. On the Geographic Allocation of Open Source Software Activities. *International Journal of Innovation in the Digital Economy* 4 (2), 25–39.
- English-Lueck, J. A., Darrach, C. N., Saveri, A., Jan. 2002. Trusting Strangers: Work Relationships in Four High-Tech Communities. *Information, Communication & Society* 5 (1), 90–108.
- Fallick, B., Fleischman, C. A., Rebitzer, J. B., Aug. 2006. Job-Hopping in Silicon Valley: Some Evidence Concerning the Microfoundations of a High-Technology Cluster. *The Review of Economics and Statistics* 88 (3), 472–481.
- Faraj, S., Jarvenpaa, S. L., Majchrzak, A., 2011. Knowledge Collaboration in Online Communities. *Organization Science* 22 (5), 1224–1239.
- Florida, R., Mellander, C., 2018. Talent, Skills, and Urban Economics. In: *The New Oxford Handbook of Economic Geography*. Oxford University Press, pp. 499–518.
- Forman, C., Goldfarb, A., Greenstein, S., Jul. 2014. Information Technology and the Distribution of Inventive Activity. *The Changing Frontier: Rethinking Science and Innovation Policy*, 169–196.
- Forman, C., Goldfarb, A., Greenstein, S., May 2016. Agglomeration of Invention in the Bay Area: Not Just ICT. *American Economic Review* 106 (5), 146–151.
- Forman, C., Goldfarb, A., Greenstein, S., 2018. How Geography Shapes - and is Shaped by - the Internet. In: *The New Oxford Handbook of Economic Geography*. Oxford University Press, pp. 269–285.
- Forman, C., Zeebroeck, N. v., Aug. 2012. From Wires to Partners: How the Internet Has Fostered R&D Collaborations Within Firms. *Management Science* 58 (8), 1549–1568.
- Gawer, A., 2009. *Platforms, Markets and Innovation: An Introduction*. In: *Platforms, Markets and Innovation*. Edward Elgar Publishing.

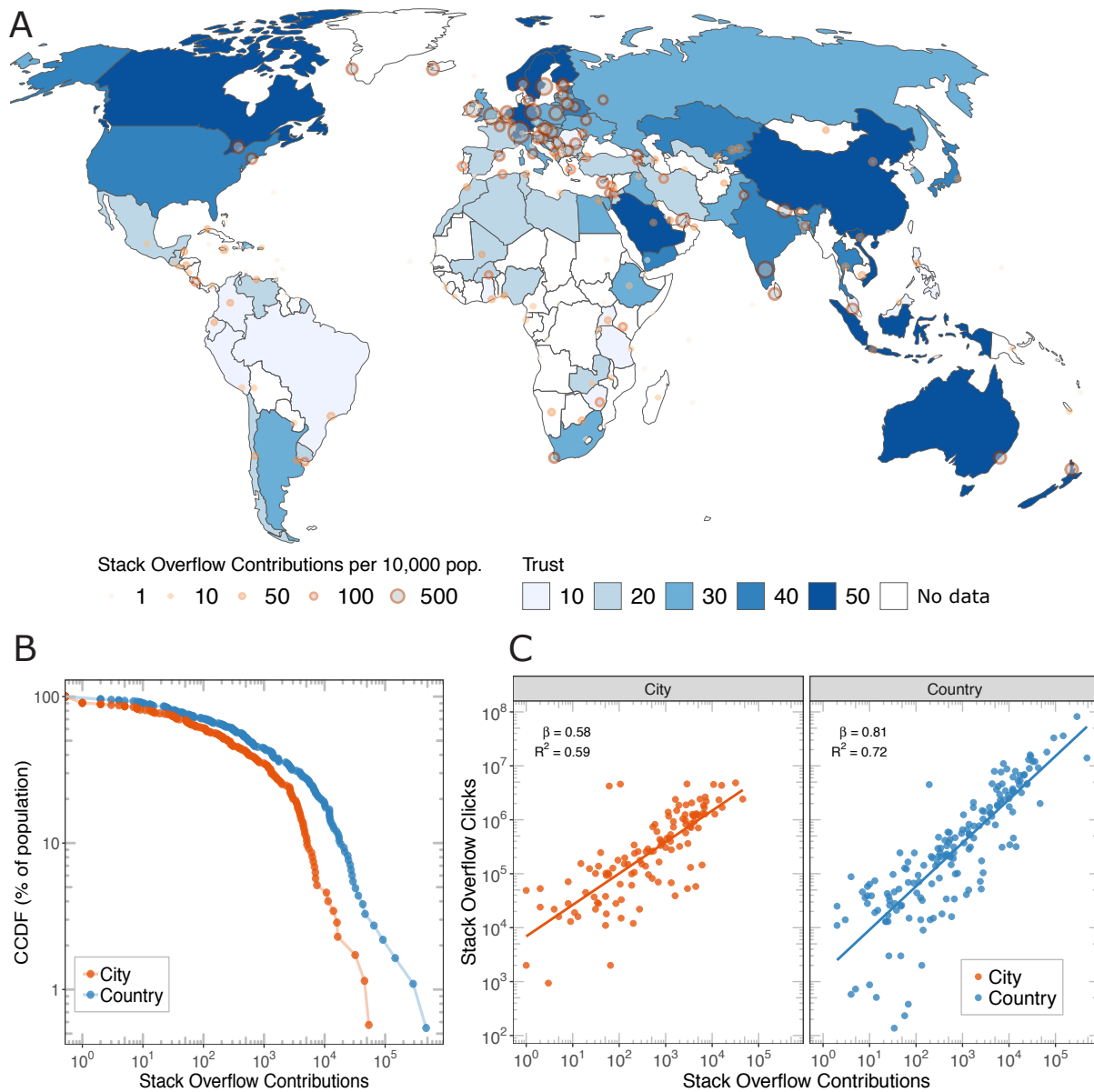
- Gertner, J., Mar. 2012. *The Idea Factory: Bell Labs and the Great Age of American Innovation*. Penguin.
- Gonzalez-Barahona, J. M., Robles, G., Andradas-Izquierdo, R., Ghosh, R. A., Dec. 2008. Geographic Origin of Libre Software Developers. *Information Economics and Policy* 20 (4), 356–363.
- Graham, M., Straumann, R. K., Hogan, B., Nov. 2015. Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia. *Annals of the Association of American Geographers* 105 (6), 1158–1178.
- Granovetter, M., Nov. 1985. Economic Action and Social Structure: The Problem of Embeddedness. *American Journal of Sociology* 91 (3), 481–510.
- Hauptman, O., Sep. 2003. Platform Leadership: How Intel, Microsoft, and Cisco Drive Industry Innovation. *Innovation* 5 (1), 91–94.
- Kobayashi, T., Ikeda, K., Miyata, K., Oct. 2006. Social Capital Online: Collective Use of the Internet and Reciprocity as Lubricants of Democracy. *Information, Communication & Society* 9 (5), 582–611.
- Lima, A., Rossi, L., Musolesi, M., Jul. 2014. Coding Together at Scale: GitHub as a Collaborative Social Network. Eighth International AAAI Conference on Weblogs and Social Media.
- Lin, J., 2011. Technological Adaptation, Cities, and New Work. *The Review of Economics and Statistics*, 21.
- Marshall, A., Jan. 2009. *Principles of Economics: Unabridged Eighth Edition*. Cosimo, Inc.
- McLuhan, M., 1964. *Understanding Media: Extensions of Man*. Routledge and Kegan Paul, London.
- Moy, C. L., Locke, J. R., Coppola, B. P., McNeil, A. J., 2010. Improving Science Education and Understanding Through Editing Wikipedia. *Journal of Chemical Education* 87 (11), 1159–1162.
- Nathan, M., Jan. 2015. Same Difference? Minority Ethnic Inventors, Diversity and Innovation in the UK. *Journal of Economic Geography* 15 (1), 129–168.
- O’Neil, M., Aug. 2014. Hacking Weber: Legitimacy, Critique, and Trust in Peer Production. *Information, Communication & Society* 17 (7), 872–888.
- Ozgen, C., Nijkamp, P., Poot, J., 2012. Immigration and Innovation in European Regions, 34.
- Penoyer, S., Reynolds, B., Marshall, B., Cardon, P. W., 2018. Impact of Users’ Motivation on Gamified Crowdsourcing Systems: A Case of Stack Overflow. *Issues in Information Systems* 19 (2).
- Sawhney, H., 1996. Information Superhighway: Metaphors as Midwives. *Media, Culture & Society* 18 (2), 291–314.
- Sobel, J., Mar. 2002. Can We Trust Social Capital? *Journal of Economic Literature* 40 (1), 139–154.
- Stephany, F., 2017. Who are your joneses? socio-specific income inequality and trust. *Social indicators research* 134 (3), 877–898.
- Stephany, F., 2019. It deepens like a coastal shelf: Educational mobility and social capital in germany. *Social Indicators Research* 142 (2), 855–885.
- Stephany, F., Braesemann, F., 2017. An Exploration of Wikipedia Data as a Measure of Regional Knowledge Distribution. In: Ciampaglia, G. L., Mashhadi, A., Yasseri, T. (Eds.), *Social Informatics*. Springer International Publishing, Cham, pp. 31–40.
- Stephany, F., Braesemann, F., et al., 2017. Bonds and bridges, and between: An empirical analysis of group-based trust. Tech. rep.
- Takhteyev, Y., Hiltz, A., 2014. Investigating the Geography of Open Source Software through Github. UToronto - Working Paper Series, 10.
- Tjoa, A. M., Tjoa, S., 2016. The Role of ICT to Achieve the UN Sustainable Development Goals (SDG). In: Mata, F. J., Pont, A. (Eds.), *ICT for Promoting Human Development and Protecting the Environment*. IFIP Advances in Information and Communication Technology. Springer International Publishing, pp. 3–13.
- Vasilescu, B., Filkov, V., Serebrenik, A., 2013. Stack Overflow and GitHub: Associations Between Software Development and Crowdsourced Knowledge. In: 2013 International Conference on Social Computing. IEEE, pp. 188–195.
- Westlund, H., Adam, F., Jun. 2010. Social Capital and Economic Performance: A Meta-analysis of 65

<u>Authors</u>	Forman and Zeebroeck (2012)	Forman et al. (2014)	Ozgen et al. (2012)	Nathan (2015)	Akomak, Weel (2009)	Davidson and Poor (2018)	Graham et al. (2015)	Engelhardt et al. (2013)
<u>Dep. Variable:</u>	<i>PATENTS</i>					<i>DIGITAL OUTCOME</i>		
	Collab. patents	Patents	Patent intensity	Indiv. patents	Patent intensity	Funded projects	Wikipedia edits	Open source programming
<i>Economics, Infrastructure</i>								
Internet	✓	✓					✓	✓
GDP/Income	✓	✓	✓			✓	✓	✓
Employment	✓	✓	✓	✓	✓			
<i>Demographics</i>								
Population		✓	✓	✓			✓	
Age		✓						
<i>Education, R&amp;D</i>								
R&D spending	✓							
Education		✓		✓	✓		✓	
Creative class						✓		
Patent stock	✓	✓				✓		
<i>Diversity and Openness</i>								
Diversity			✓	✓				
Ethnic groups		✓	✓					
Trust			✓		✓			
<i>Study Details</i>								
Geography	US	US	Europe	UK	Global	US	Global	Global
N	18,860	2,793	340	210,008	102	2,743	160	160
Unit	Patents	Patents	Regions	Patents	Countries	Projects	Countries	Countries
Source	Table 9	Table 3	Table 5	Table 8	Table 6	Table 2	Table 5	Figures 9-10

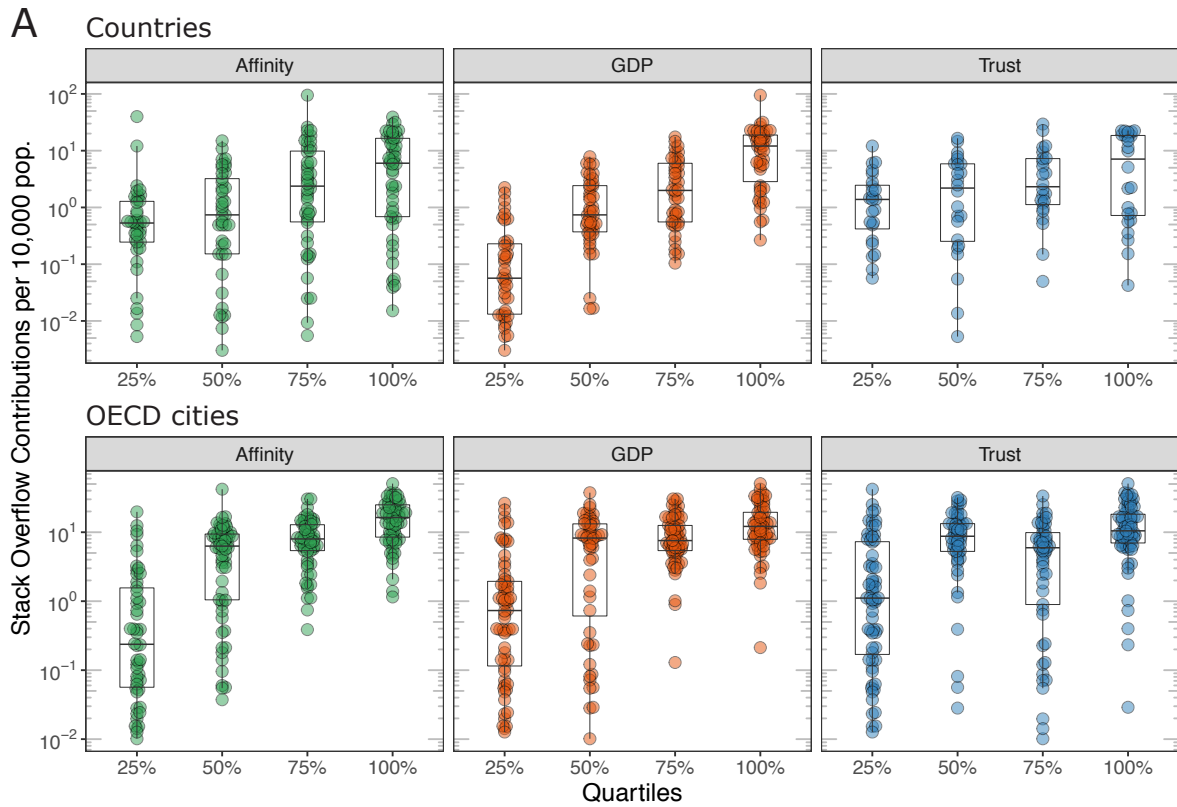
**Tab. 1.** Summary of relevant empirical studies on the geography of 'traditional' and 'digital' forms of innovative activity: In investigations with a focus on patents, economic controls, such as employment, income or GDP, together with measures of population density and education, are used to model the agglomeration of innovative activity. In studies focusing on digital innovation outcomes, broadband connectivity and GDP are most commonly used to explain the geographic distribution. Measures on societal openness or trust have, so far, not been included in studies on the geography of digital platform contributions.

Studies. European Planning Studies 18 (6), 893–919.

Yoo, Y., Boland, R. J., Lyytinen, K., Majchrzak, A., Oct. 2012. Organizing for Innovation in the Digitized World. Organization Science 23 (5), 1398–1408.



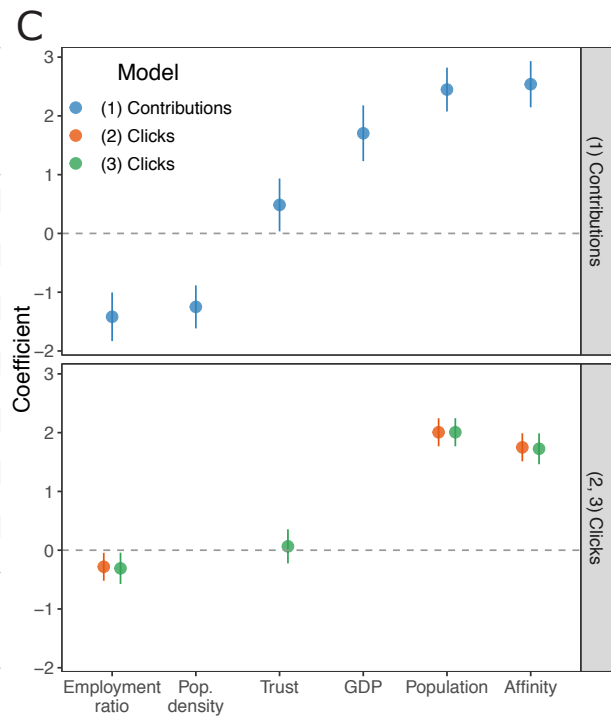
**Fig. 1.** (A) National levels of *trust* (country-colour) and Stack Overflow contributions per city (circle size; the most active city per country is shown): Cities with higher per capita contributions tend to be located in countries with relatively high levels of societal trust. (B) Complementary cumulative distribution functions of country- and city-level contributions (log-log scale): The contributions are highly concentrated and follow a fat-tailed distribution - the most active eight cities (5%) in the dataset account for half of the total number of contributions. (C) Comparison of country- and city-level contributions and clicks to Stack Overflow: Digital knowledge consumption (clicks) and production (contributions) are highly correlated, but not identical.



**B**

Dependent variable (OLS; IHS-transf.)	Stack Overflow		
	Contributions (optimised) (1)	Clicks (optimised) (2)	Clicks (incl. trust) (3)
<b>Affinity</b>	2.03***	1.40***	1.38***
(city, log scale)	(0.16)	(0.10)	(0.11)
<b>Population</b>	1.54***	1.26***	1.26***
(city, log scale)	(0.12)	(0.08)	(0.08)
<b>Pop. density</b>	-0.65***		
(city, log scale)	(0.10)		
<b>GDP</b>	1.89***		
(city, log scale)	(0.27)		
<b>Employment ratio</b>	-5.24***	-1.04**	-1.14**
(city, log scale)	(0.78)	(0.45)	(0.50)
<b>Trust</b>	1.72**		0.23
(region)	(0.81)		(0.52)
<b>Constant</b>	-9.33***	-1.25	-0.91
	(3.55)	(2.19)	(2.32)
Observations	266	266	266
R <sup>2</sup>	0.76	0.68	0.68
Adjusted R <sup>2</sup>	0.75	0.67	0.67

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01



**Fig. 2. (A)** Country- and OECD city-level distributions of Stack Overflow contributions by internet affinity, GDP, and Trust (grouped in quartiles): all three variables are positively associated with the number of contributions per capita. **(B)** Regression results on the determinants of contributions and clicks to Stack Overflow: While contributions and clicks positively relate to affinity and population size, only contributions are affected by trust. **(C)** Dot-Whisker plot of regression models with normalised coefficients (coefficients divided by two standard deviations).

## Appendix

Statistic - per city <sup>*</sup>	N	Mean	Min	Median	Max
<u>Stack Overflow</u>					
Contributions (city level)					
<i>Untransformed</i> (22 missing)	244	2,178	1	748	45,634
<i>Untransformed</i> (missing values set to zero)	266	1,998	0	637	45,634
<i>IHS-transf.</i>	266	6.31	0	7.15	11.42
Clicks (per unique IP-address)					
<i>Clicks</i> (city level)	266	429,212	778	178,000	4,834,000
User affinity <sup>a)</sup>					
<i>Affinity</i> (city level)	266	1.03	0.22	0.82	4.78
<u>Trust</u>					
Share of pop. having generalized trust <sup>b)</sup>					
<i>Trust</i> (regional level)	266	0.35	0	0.37	0.80
<u>Internet &amp; Economy</u>					
Population with broadband access (%)					
<i>Broadband</i> (regional level)	266	74.13	18	76.9	100
GDP per capita (2010 US-Dollars)					
<i>GDP</i> (city level)	266	39,494	7,315	37,932	86,730
Employment rate <sup>c)</sup>					
<i>Employment rate</i> (city level)	266	68.67	37.63	69.49	88.07
<u>Demographics</u>					
Population size					
<i>Population</i> (city level)	266	2,049,278	436,627	1,031,690	35,925,101
Population density per square mile					
<i>Pop. density</i> (city level)	266	721.79	33.74	495.50	5,339
Age-Dependency-Ratio <sup>d)</sup>					
<i>Age-dep. ratio</i> (city level)	266	23.84	6.08	22.35	50.94
Pop. with at least secondary education (%)					
<i>Education</i> (regional level)	266	79.75	27.90	85.20	97.70

a) A statistical value, based on the global distribution of visitors of Stack Overflow. For example, San Francisco has an affinity value of 1.94, that means a given visitor to Stack Overflow is twice as likely to be from San Francisco as any Internet user chosen at random. (Source: <https://www.quantcast.com/measure/stackoverflow.com?country=US#/generalInterestsCard>)

b) The generalized trust question reads "Generally speaking, do you think that most people can be trusted (or do you think that you can never be to careful in dealing with people)?" In some cases, respondents are asked to answer on a Likert-Scale from 0 (no trust) to 10 (full trust). In this case, individuals, who replied with a trust level of six and higher are counted.

c) Employment rates are defined as a measure of the extent to which available labour resources (people available to work) are being used. They are calculated as the ratio of the employed to the working age population.

d) Age dependency ratio: number of individuals older than 65 years, divided by the number of individuals between 15 and 65 years.

<sup>\*</sup>) The 266 metropolitan areas are: (AUS) Adelaide, Brisbane, Gold Coast City, Melbourne, Perth, Sydney, (AUT) Graz, Linz, Vienna, (BEL) Antwerp, Brussels, Ghent, Lige, (CAN) Calgary, Edmonton, Hamilton, Montreal, Ottawa, Quebec, Toronto, Vancouver, Winnipeg, (CHE) Basel, Geneva, Zurich, (CHL) Concepcion, Santiago, Valparaiso, (CZE) Brno, Ostrava, Prague, (DEU) Aachen, Augsburg, Berlin, Bochum, Bonn, Bremen, Cologne, Dortmund, Dresden, Duisburg, Dsseldorf, Essen, Frankfurt, Freiburg, Hamburg, Hanover, Karlsruhe, Leipzig, Mannheim, Munich, Mnster, Nuremberg, Saarbrücken, Stuttgart, (DNK) Copenhagen, (ESP) Barcelona, Bilbao, Las Palmas de Gran Canaria, Madrid, Mlaga, Seville, Valencia, Zaragoza, (EST) Tallinn, (FIN) Helsinki, (FRA) Bordeaux, Grenoble, Lille, Lyon, Marseille, Montpellier, Nantes, Nice, Paris, Rennes, Rouen, Saint-tienne, Strasbourg, Toulon, Toulouse, (GBR) Birmingham, Bradford, Bristol, Cardiff, Edinburgh, Glasgow, Leeds, Leicester, Liverpool, London, Manchester, Newcastle upon Tyne, Nottingham, Portsmouth, Sheffield, (GRE) Athens, Thessaloniki, (HUN) Budapest, (IRE) Dublin, (ITA) Bari, Bologna, Catania, Florence, Genova, Milan, Naples, Palermo, Rome, Turin, Venice, (JPN) Anjo, Fukuoka, Fukuyama, Hamamatsu, Himeji, Hiroshima, Kagoshima, Kanazawa, Kitakyushu, Kochi, Kofu, Kumamoto, Kurashiki, Maebashi, Matsuyama, Mito, Nagano, Nagasaki, Nagoya, Naha, Niigata, Numazu, Oita, Okayama, Osaka, Sapporo, Sendai, Shizuoka, Takamatsu, Tokushima, Tokyo, Toyama, Toyohashi, Utsunomiya, Wakayama, Yokkaichi, (KOR) Busan, Changwon, Cheongju, Daegu, Daejeon, Gwangju, Jeonju, Pohang, Seoul, Ulsan, (MEX) Acapulco, Aguascalientes, Celaya, Chihuahua, Ciudad Juarez, Cuernavaca, Culiacn, Durango, Guadalajara, Hermosillo, Heroica Veracruz, Irapuato, Len, Mrida, Mexicali, Mexico City, Monterrey, Morelia, Oaxaca, Pachuca, Puebla City, Reynosa, Saltillo, San Luis Potosi, Santiago de Quertaro, Tampico, Tijuana, Torren, Tuxtla Gutierrez, Xalapa, (NLD) Amsterdam, Eindhoven, Rotterdam, The Hague, Utrecht, (NOR) Oslo, (POL) Gdask, Katowice, Krakw, d, Lublin, Pozna, Warsaw, Wrocow, (POR) Lisbon, Porto, (SVK) Bratislava, (SVN) Ljubljana, (SWE) Gothenburg, Malm, Stockholm, (USA) Akron, Albany, Albuquerque, Atlanta, Austin, Baltimore, Baton Rouge, Birmingham, Boston, Buffalo, Charleston, Charlotte, Chicago, Cincinnati, Cleveland, Colorado Springs, Columbia, Columbus, Dallas, Dayton, Denver, Des Moines, Detroit, El Paso, Fort Worth, Fresno, Grand Rapids, Harrisburg, Houston, Indianapolis, Jacksonville, Kansas City, Las Vegas, Little Rock, Los Angeles, Louisville, Madison, McAllen, Memphis, Miami, Milwaukee, Minneapolis, Nashville, New Orleans, New York, Norfolk, Oklahoma City, Omaha, Orlando, Philadelphia, Phoenix, Pittsburgh, Portland, Providence, Raleigh, Richmond, Sacramento, Saint Petersburg, Salt Lake City, San Antonio, San Diego, San Francisco, Seattle, St. Louis, Tampa, Toledo, Tucson, Tulsa, Washington, Wichita.

Tab. 2. Summary of outcome and control variables.



A

Step:	Dependent variable: Stack Overflow contributions (IHS-transf.)								
	FW 1 (1)	FW 2 (2)	FW 3 (3)	FW 4 (4)	FW 5 (5)	OPTIMAL (6)	BW 3 (7)	BW 2 (8)	BW 1 (9)
<i>Affinity</i> (city, log scale)	2.96*** (0.21)	2.72*** (0.17)	2.56*** (0.16)	2.26*** (0.16)	2.14*** (0.15)	2.03*** (0.16)	2.02*** (0.16)	1.99*** (0.17)	2.00*** (0.178)
<i>Population</i> (city, log scale)		1.61*** (0.14)	1.80*** (0.13)	1.63*** (0.13)	1.50*** (0.12)	1.54*** (0.12)	1.52*** (0.12)	1.52*** (0.12)	1.54*** (0.13)
<i>Pop. density</i> (city, log scale)			-0.77*** (0.11)	-0.67*** (0.10)	-0.61*** (0.10)	-0.65*** (0.10)	-0.64*** (0.10)	-0.65*** (0.10)	-0.67*** (0.11)
<i>GDP</i> (city, log scale)				1.28*** (0.23)	2.10*** (0.25)	1.89*** (0.27)	2.15*** (0.35)	2.10*** (0.37)	2.11*** (0.37)
<i>Employment ratio</i> (city, log scale)					-4.80*** (0.75)	-5.24*** (0.78)	-5.11*** (0.78)	-5.01*** (0.82)	-5.04*** (0.82)
<i>Trust</i> (region)						1.72** (0.81)	1.84** (0.81)	1.78** (0.83)	1.66* (0.87)
<i>Education</i> (region, log scale)							-0.62 (0.54)	-0.74 (0.61)	-0.86 (0.66)
<i>Broadband</i> (region, log scale)								0.25 (0.58)	0.27 (0.58)
<i>Age-dep. ratio</i> (city, log scale)									0.13 (0.29)
<i>Constant</i>	6.80*** (0.14)	-15.92*** (1.94)	-13.88*** (1.79)	-25.54*** (2.72)	-12.39*** (3.27)	-9.33*** (3.55)	-9.70*** (3.57)	-10.11*** (3.69)	-10.10*** (3.70)
Observations	266	266	266	266	266	266	266	266	266
R <sup>2</sup>	0.42	0.62	0.68	0.72	0.76	0.76	0.76	0.76	0.76
Adjusted R <sup>2</sup>	0.42	0.62	0.68	0.71	0.75	0.75	0.75	0.75	0.75

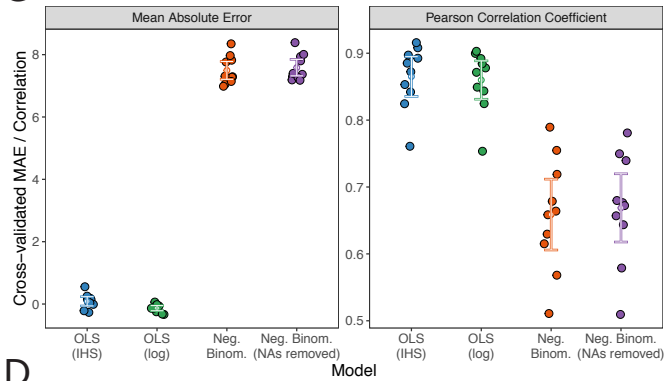
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

B

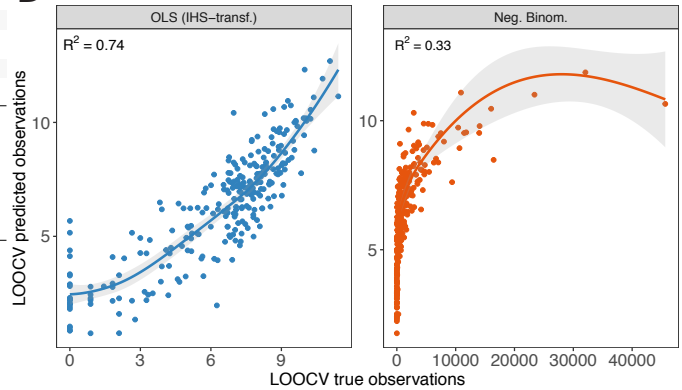
	Dependent variable: Stack Overflow contributions			
	Continuous		Discrete	
	OLS IHS-transf.	log-transf.	Negative binomial NAs repl. by 0	Negative binomial NAs removed
	(1)	(2)	(3)	(4)
<i>Affinity</i> (city, log scale)	2.03*** (0.16)	1.58*** (0.14)	1.57*** (0.13)	1.19*** (0.11)
<i>Population</i> (city, log scale)	1.54*** (0.12)	1.15*** (0.10)	1.28*** (0.10)	1.10*** (0.08)
<i>Pop. density</i> (city, log scale)	-0.65*** (0.10)	-0.33*** (0.08)	-0.32*** (0.08)	-0.18*** (0.07)
<i>GDP</i> (city, log scale)	1.89*** (0.27)	2.08*** (0.21)	1.63*** (0.22)	1.66*** (0.17)
<i>Employment ratio</i> (city, log scale)	-5.24*** (0.78)	-3.49*** (0.63)	-3.57*** (0.63)	-2.04*** (0.52)
<i>Trust</i> (region)	1.72** (0.81)	1.61** (0.64)	1.38** (0.65)	1.41*** (0.53)
<i>Constant</i>	-9.33*** (3.55)	-15.69*** (2.82)	-11.00*** (2.85)	-17.11*** (2.33)
Observations	266	244	266	244
R <sup>2</sup>	0.76	0.77		
Adjusted R <sup>2</sup>	0.75	0.76		
Log Likelihood			-1,954	-1,869
$\theta$			0.78*** (0.07)	1.22*** (0.10)
Akaike Inf. Crit.			3,922	3,752

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

C



D



**Fig. 3. (A)** Step-wise forward and backward model selection: Both selection procedures lead to the same optimal model in which OECD city-level contributions are explained by affinity, population, population density, GDP, employment rate, and trust. **(B)** Regression models for continuous (OLS) and discrete (Negative Binomial) outcome measures: Models (2) and (4) exclude cities without contributions. The two continuous and discrete outcome models lead to comparable results. **(C)** Ten-fold cross-validation of the four models: The linear regression models outperform the negative binomial models in terms of out-of-sample prediction accuracy (mean absolute error and pearson correlation) **(D)** Leave-one-out cross-validation of models (1) and (3): While the OLS model predicts most test-observations well ( $R^2 = 0.74$ ), the negative binomial model fails to accurately predict large values.