Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-020-00685-7 OPEN

Predicting urban innovation from the US Workforce Mobility Network

Moreno Bonaventura[1], Luca Maria Aiello[2], Daniele Quercia[2,3]✉ & Vito Latora[1,4,5]

While great emphasis has been placed on the role of social interactions as a driver of innovation growth, very few empirical studies have explicitly investigated the impact of social network structures on the innovation performance of cities. Past research has mostly explored scaling laws of socio-economic outputs of cities as determined by, for example, the single predictor of population. Here, by drawing on a publicly available dataset of the startup ecosystem, we build the first Workforce Mobility Network among metropolitan areas in the US. We found that node centrality computed on this network accounts for most of the variability observed in cities' innovation performance and significantly outperforms other predictors such as population size or density, suggesting that policies and initiatives aiming at sustaining innovation processes might benefit from fostering professional networks alongside other economic or systemic incentives. As opposed to previous approaches powered by census data, our model can be updated in real-time upon open databases, opening up new opportunities both for researchers in a variety of disciplines to study urban economies in new ways, and for practitioners to design tools for monitoring such economies in real-time.

[1] School of Mathematical Sciences, Queen Mary University of London, London E14NS, UK. [2] Nokia Bell Labs, Cambridge CB30FA, UK. [3] CUSP, King's College London, London WC2R2LS, UK. [4] Dipartimento di Fisica e Astronomia, Università di Catania and INFN, 95123 Catania, Italy. [5] The Alan Turing Institute—The British Library, London NW12DB, UK. ✉email: quercia@cantab.net

Introduction

Over the last 2 decades, developed and developing countries alike have witnessed a radical transformation in the nature and dynamics of their innovation processes. A major factor that has triggered this change is the emergence of new entrepreneurial ecosystems centered on high-growth startups. In the United States, startups account for the majority of new job creations (Decker et al., 2014) and have rapidly expanded not only in size but also geographically by creating distributed innovation centers (Acs and Mueller, 2008). Abundant empirical evidence supports the idea that young and innovative firms guarantee the long-term growth of cities and sustain the economic life by creating wealth and new jobs also in related industries (Bos and Stam, 2014; Glaeser et al., 2010; Hall and Raumplaner, 1998, Haltiwanger et al., 2013; Mumford, 1961; Weins and Jackson, 2014).

Researchers have tried to shed light on early indicators of success in modern innovation environments. In the attempt of building baseline models to predict innovation in cities, past efforts have mainly focused on predicting a wide range of socio-economic indicators of wealth (e.g., GDP, employment, housing and infrastructures) and a range of *innovation* indicators (e.g., abundance of young firms, number of patents granted) solely based on population size or density (Arbesman et al., 2009; Arcaute et al., 2015; Bettencourt et al., 2007a, b). These studies have shown that population size alone is able to reliably predict—with a coefficient of determination R^2 for linear regression in the [0.88, 0.99] range—several socio-economic outputs of cities including income, electrical consumption, total wages, and employment. Yet, the correlations between population characteristics and outputs associated with *innovation* processes such as number of granted patents ($R^2 = 0.72$), number of inventors ($R^2 = 0.76$), and R&D establishments ($R^2 = 0.77$) are not equally strong. In fact, innovation-related indicators report the smallest correlation coefficients among all the other variables (Bettencourt et al., 2007a) (Fig. 1).

This discrepancy points to three main limitations of prediction models solely based on demographic variables. First, by treating geographical areas as isolated entities, such models overlook the role of social interactions, yet well-established urban theories (Jacobs, 1970) and qualitative (Saxenian, 1996) and quantitative findings in economics (Glaeser, 2011) have repeatedly shown that a dense and dynamic web of interactions among specialized workers, entrepreneurs, and investors—also referred to as the "thickness of the market"—plays a pivotal role in driving idea recombination, innovation generation, and ultimately economic growth (Glaeser and Scheinkman, 2001; Jacobs, 1961; Moretti, 2012). Second, these past models do not account for the fact that cities grow through the attraction of highly talented individuals (also called "the creative class" (Florida, 2005)), and the creative outputs from such individuals have been recently found to explain superlinear urban scaling (Keuschnigg et al., 2019). Finally, the life-cycle of a modern innovative startup—its birth, growth, acquisition, and extinction—is much faster than the time frames within which past models' inputs (e.g., demographic data) and outputs (e.g., patenting rates) are typically defined.

Previous research has provided evidence that simple scaling laws of population miss evolutionary dynamics that are key to explain many city-level processes (Depersin and Barthelemy, 2018), and that the application of tools from statistical physics to a variety of spatial networks allows for a more accurate description of such complex dynamics (Barbosa et al., 2018; Barthelemy, 2016, 2019; Kirkley et al., 2018; Lämmer et al., 2006; Tria, 2014). However, constrained by limited data availability, only a few empirical studies have attempted to investigate the impact of different types of social network structures on economic growth and innovation performance of cities (Bettencourt et al., 2007; Eagle et al., 2010; Makarem, 2016, Powell et al., 1996; Sorenson and Stuart, 2001).

This work contributes to fill the gap by drawing on a novel dataset from CrunchBase, an online database containing historical records of the evolution of the worldwide startup ecosystem. In previous research, CrunchBase records have been used to predict the success of individual startups (Moreno et al., 2020). Our research question is: To which extent proxies for the US workforce mobility inferred from CrunchBase predict two main urban innovation metrics? To answer that question, we built and analyzed the first *Workforce Mobility Network* (WMN), which, unlike previous approaches in the literature, is temporally fine-grained and comes from publicly available data[1]. The network's nodes are metropolitan areas, and its directed links (edges) are workforce flows between area pairs; the edge weight from metropolitan area i to j is equal to the number of professionals who worked at i and then moved for work to j. Figure 2 provides an illustration of the procedure adopted to construct WMN:

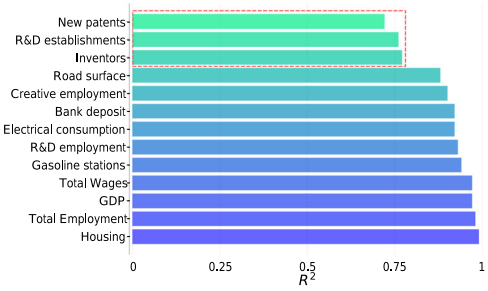

Fig. 1 Correlation coefficients between population size and various socio-economic outputs, adapted from Bettencourt et al. (2007a). Data suggest that innovation-related indicators (marked in red) are less correlated with population than other socio-economic outputs. These results are computed with data covering the [1997, 2003] period.

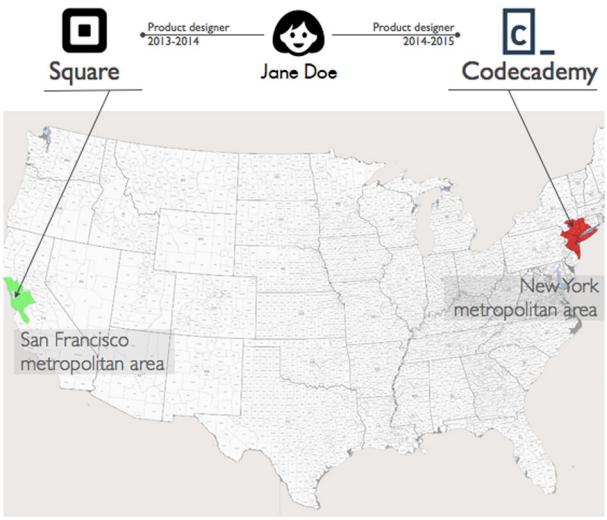

Fig. 2 Example of how the Workforce Mobility Network (WMN) is built. Its nodes represent metropolitan statistical areas (MSA) in the United States, while each of its directed links has a weight equal to the number of employees who worked in one area and then moved to another area. For instance, Jane Doe, who moved from Square Inc. in the San Francisco area (green) to Codecademy in the New York area (red), acted as a bridge between the two companies and contributed to increase the weight of the directed link from the San Francisco area to the New York area by one.

Dr. Jane Doe quits her job at Square Inc., a company based in the San Francisco–Oakland–Hayward metropolitan area (green), to then join Codecademy, located in the New York–Newark–Jersey area (red), thus acting as a bridge between the two areas; ultimately, the directed link from the "San Francisco–Oakland–Hayward" node to the "New York–Newark–Jersey" node has a weight equal to the number of unique workers who moved from one location to the other.

The opportunity to recombine ideas and access relevant knowledge is crucial for companies that aim at generating innovation (Burt, 1993; Hargadon, 1998; Parise et al., 2015). The likelihood of a company benefiting from new ideas, know-how, and talents is determined not only by the availability of these resources within the city where the company is located (endogenous view suggested by research on urban complexity (Bettencourt et al., 2007; Eagle et al., 2010; Makarem, 2016; Powell et al., 1996; Sorenson and Stuart, 2001)), but also by the opportunity to absorb them from other cities (exogenous view suggested by research on the economics of migration (Florida, 2005; Glaeser, 2011; Keuschnigg et al., 2019)). As such, we hypothesized that the *most central areas* in WMN, rather than the most densely populated ones, are the most innovative. In so doing, we do not study what determines migration: it is known that workforce mobility impacts network centrality (opportunities are created by talent migration (Keuschnigg et al., 2019)), and that, in a circular way, network centrality impacts workforce mobility (talent migration happens where opportunities are (Florida, 2005)). Instead, we study to which extent network centrality metrics are predictive of economic performance. To that end, we considered two innovation measures for each metropolitan area i: (1) the number $\mathcal{S}_i$ of successful startups in i (a startup is successful if it either was acquired, did an Initial Public Offering (IPO), or acquired another startup); and (2) the cumulative acquisition price $\mathcal{A}_i$ of all startups in i. Differently from commonly used measures of output such as the number of granted patents, our measures adapt more dynamically to the rapidly changing market and better reflect a startup's ability to translate its innovation potential into immediate and tangible economic value. In a modern innovation landscape characterized more and more by digital solutions, global outreach, low barrier to entry, and extremely fast business developments, the number of patents might not fully reflect actual levels of innovation. Often patents are used as a defensive tool against "patent trolling" (Cohen et al., 2016) or are used to discourage the entry of market newcomers rather than actually being used to produce and commercialize genuinely innovative products (Nicholas, 2013). For completeness, we present empirical results considering patenting rates as a proxy for innovation as well, and do so in *Supplementary Information*.

In summary, we measured to which extent WMN—specifically, the centrality of its nodes—predicts innovation performance of cities, measured through $\mathcal{S}_i$ and $\mathcal{A}_i$, and how those predictions compare to previous models' in the literature.

Methods

Datasets. We combined data from three sources. First, from the 2010 US census data, we extracted information about population size, land area, and population density at the level of *Metropolitan Statistical Area* (MSA). Second, from the United States Patent and Trademark Office (USPTO), we associated the numbers of patents granted in the year of 2010 with the inventors' metropolitan areas. Third, from the CrunchBase web APIs, we collected all information regarding *organizations* recorded up to the end of 2016, and for *people* (workers) recorded up to end of year 2010. For each organization we extracted data on: address of the headquarter, foundation date, funding rounds, acquisitions (also referred to as *exits*), initial public offers (IPOs), status (active, closed), and team members. The address, in turn, consists of street name, zip-code, city name, and state. Funding rounds record the financial investment of individuals or venture capital firms into a company (organization), i.e., the purchase of a certain percentage of ownership of the company, while acquisitions indicate the transfer of the company's total ownership to another company. The data on funding rounds and acquisitions include the parties involved, the date, and the monetary value of the transaction in US dollars. We were able to associate the companies in our data with 369 (out of the 374) metropolitan areas. Workers are linked to organizations through the professional roles they hold. Examples of role titles are *CEO*, *founder*, *board member*, and *employee*. Workers can have multiple jobs/roles within the same organization or across different organizations. Roles can be associated with a start date and an end date; the earliest starting dates in the dataset are in the year of 1960, but 75% of the records are from 2000 to 2010 (see *Supplementary Information*). About 42% of all the job records include a starting date allowing for a longitudinal analysis of the flow of workers between various firms.

Construction of the Workforce Mobility Network. We modeled the Workforce Mobility Network (WMN) as a directed graph of metropolitan areas. Given any pair of roles r_1 and r_2 played by a worker in metropolitan areas i and j, respectively ($i \neq j$), we incremented the weight w_{ij} by one if the *start date* in role r_1 preceded the *start date* in role r_2. When *end dates* were available, we incremented both weights w_{ij} and w_{ji} by one if the *end date* of r_1 followed the *start date* of r_2—in that case, the roles temporally overlapped and we, therefore, assumed that information exchange between the two areas was bidirectional.

Centrality measures. Different measures of centrality have been proposed over the years to quantify the importance of a node in a complex network (Latora et al., 2017). In this work, we computed four centrality measures for each WMN node: degree centrality, node strength, harmonic closeness, and Google PageRank.

Let G be a weighted graph with N nodes described by the $N \times N$ weighted adjacency matrix $W = \{w_{ij}\}$ whose entry w_{ij} is equal to the weight of the directed link connecting node i to node j, or is equal to 0 if there is not a direct connection from node i to node j. As for the case of G being an unweighted graph, we define the adjacency matrix $A = \{a_{ij}\}$ of G, which simply indicates which pairs of nodes are connected with a $N \times N$ matrix such that $a_{ij} = 1$ if $w_{ij} \neq 0$, and $a_{ij} = 0$ if $w_{ij} = 0$.

Our first centrality measure out of the four is *degree centrality*, which is based on the idea that important nodes are those with the largest number of ties to other nodes in the graph. In a directed graph, the degree centrality of node i is defined as:

$$C_i^D = \frac{k_i}{N-1} = \frac{\sum_{j=1}^{N} a_{ij} + a_{ji}}{2(N-1)} \quad (1)$$

where k_i is the number of directed links to i and those from i.

Our second centrality measure is *strength centrality*. For each node i, this is defined as:

$$C_i^S = \frac{s_i}{\sum_j s_j} = \frac{\sum_{j=1}^{N} w_{ij} + w_{ij}}{\sum_{i,j} w_{ij}} \quad (2)$$

where strength s_i of node i is the sum of the weights of the edges incident in i.

Our third centrality measure is the harmonic closeness centrality (Marchiori and Latora, 2000). For each node i, this measure is the sum of all the minimum distances d_{ij} from i to any

another node j. The minimum distance d_{ij} is the length of the weighted shortest path between i and j, considering that the distance between two neighbors a and b is equal to the inverse of the edge weight that connects them ($d_{ab} = \frac{1}{w_{ab}}$). Formally, the harmonic centrality is defined as:

$$C_i^C = \sum_j \frac{1}{d_{ij}}. \quad (3)$$

Our fourth and final centrality measure is the *PageRank centrality*. For each node i, this is the stationary probability that a "surfer" that randomly travels on the network's directed links arrives at node i. It is recursively defined as:

$$C_i^{PR} = \frac{1-\alpha}{N} + \alpha \cdot \sum_{\{j|(i,j)\in A\}} \frac{C_j^{PR}}{k_j} \quad (4)$$

where k_j is the degree of node j, and α is a *damping factor* (traditionally set to 0.85) that models the probability of the surfer following an existing directed link instead of jumping to any other node picked at random with uniform probability. In this work, we considered a weighted version (Xing and Ghorbani, 2004) of the PageRank centrality that sets the probability of following a directed link proportional to the weight of that link. Formally, this is expressed as:

$$C_i^{PR} = \frac{1-\alpha}{N} + \alpha \cdot \sum_{\{j|(i,j)\in A\}} C_j^{PR} \cdot \frac{w_{ji}}{s_j} \quad (5)$$

where the factor $\frac{w_{ji}}{s_j}$ expresses the probability of transitioning from node j to node i being equal to the weight of the link between j and i (w_{ji}) divided by the total strength of j's outgoing links (s_j). The PageRank values are computed with an iterative procedure (implemented efficiently through the so-called power method (Arasu et al., 2002)) that starts by assigning a uniform PageRank value to all nodes $C_i^{PR} = 1/N$, and runs until convergence.

For all the four centrality measures, we considered their normalized versions $\hat{C}_i = \frac{C_i}{\sum_{j=1}^{N} C_j}$ such that the sum of centrality scores over all the nodes in the network is equal to 1.

Results

All the following models are based on startups that were active in the United States in 2010, and on all their historical information up to the end of that year. For each of the metropolitan areas in which these startups were located, we measured the innovation performance indicators $\mathcal{S}_i$ and $\mathcal{A}_i$ in the [2011–2016] period.

Residual variability of population-based models. Consistently with previous work (Bettencourt et al., 2007a), we found a non-linear scaling of our two innovation measures $\mathcal{S}_i$ and $\mathcal{A}_i$ with population size $\mathcal{P}_i$, and with past fundings $\mathcal{F}_i$ (Fig. 3): the two innovation measures scale superlinearly with population size ($\beta \approx 1.2 - 1.6$, suggesting increasing returns with population size), and, as one expects for any material quantity, they scale sub-linearly with past fundings ($\beta \approx 0.6 - 0.8 < 1$, which "characterizes material quantities displaying economies of scale associated with infrastructure" (Bettencourt et al., 2007a)).

However, despite the correlations being strong (the cross-correlations are shown in *Supplementary Information*), performance variability is still high. Many cities that are similar in size and in past fundings expressed very different performances. For example, the *North Port-Bradenton-Sarasota* metropolitan area (Florida) and the *Colorado-Springs* metropolitan area (Colorado) are very similar with respect to number of startups active in 2010 (respectively, 106 and 99), population (~10^6), and funding received (~10^8\$), yet the performances of their companies are significantly different: companies in "North Port-Bradenton-Sarasota" have been sold for a cumulative value of $5.8 \cdot 10^9$\$, while those in "Colorado-Springs" reported a cumulative acquisition price smaller by two orders of magnitude, namely $4.3 \cdot 10^7$\$.

Our aim was to investigate to which extent these differences in performance could be accounted for by other predictors. In particular, we hypothesized that workforce mobility explains most of the residual variability.

The Workforce Mobility Network. We constructed the Workforce Mobility Network (WMN) among metropolitan areas by using CrunchBase records of job transitions from 1960 to the end of 2010. Among the 380 metropolitan areas in the United States, 243 had at least one active startup in our data. As a result, the final network had 243 nodes and 2,169 edges, and reflected 26,660 worker flows among metropolitan areas. When considering both incoming and outgoing edges, the maximum node degree is 165, and the maximum node strength (the maximum sum of the link weights for a node) is 8370. The strength distribution follows a power-law function with an exponent ~ 2, a value similar to those observed in other real-world weighted networks (Latora et al., 2017).

To visualize WMN, we projected it onto the map of the United States, centering its nodes on the metropolitan areas they represent (Fig. 4A). Since the number of edges was high, to improve the visualization, we reduced the number of displayed edges with a network backbone extraction algorithm (Coscia and Neffke, 2017), which identified the most statistically significant edges for each node and pruned the rest out. Then, on the original WMN (that not subject to any backboning), we computed each node's centrality according the four measures defined in *Methods*, and PageRank yielded the best fit. In Fig. 4, we notice that the most central nodes tend to be US coastal areas, which happen to be linked with each other by the strongest edges. Although population and centrality are in general well correlated (Spearman rank correlation $\rho = 0.70$), large fluctuations are still observed: indeed, despite being large, several cities do not score high in terms of node centrality (Fig. 4B).

To identify cities that are small yet central, and viceversa, we ranked cities by their ratios η between their PageRank centrality values C_i^{PR} and their population sizes $\mathcal{P}_i$:

$$\eta = \frac{C_i^{\mathrm{PR}} / \sum_j C_j^{\mathrm{PR}}}{\mathcal{P}_i / \sum_j \mathcal{P}_j}. \quad (6)$$

Both centrality values and population sizes are normalized by their sums across all areas. Table 1 shows the 10 metropolitan areas with the highest values of η, and the 10 with the lowest values. Metropolitan areas at the top have higher centrality relative to their population size. These include large and central areas such as San Francisco as well as much smaller ones (e.g., Boulder and Ithaca) that are remarkably central despite their limited size. On the other hand, the ten cities at the bottom are generally very populous yet not central in workforce flows, and, with the exception of Virginia Beach, the remaining nine cities experience relatively limited financial returns from innovation. These findings seem to suggest that network centrality might predict innovation performance better than what population counts would do. We set out to test that proposition next.

Predicting innovation performance of cities. We used linear regression to evaluate the impact of demographic characteristics and network characteristics on the performance of an area's startups. Linear regression is an approach for modeling a linear relationship between a dependent variable (our innovation measure $\mathcal{S}_i$ or $\mathcal{A}_i$) and a set of independent variables, and it does so by associating a

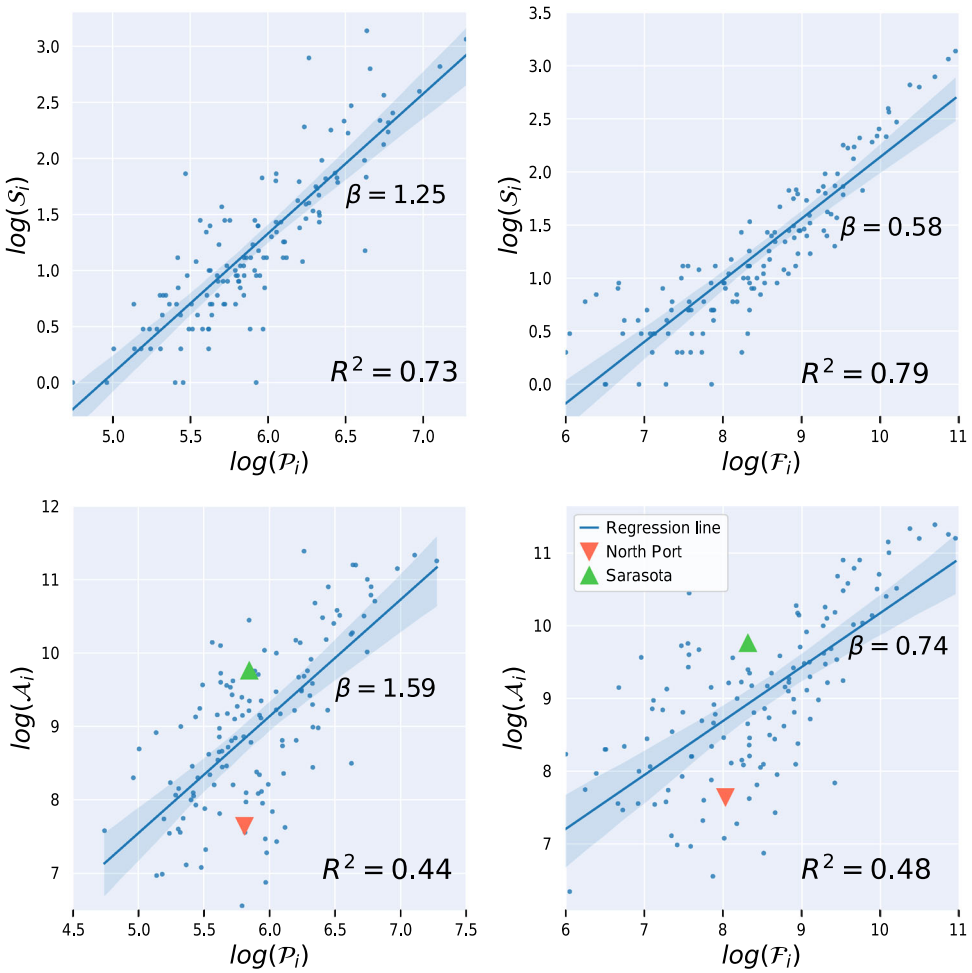

Fig. 3 Scatter plots of our two innovation measures, namely the total number of successful startups $\mathcal{S}_i$ and the cumulative acquisition price $\mathcal{A}_i$, against population size $\mathcal{P}_i$ and total past fundings $\mathcal{F}_i$. Double logarithmic plots, coefficients of determination R^2, and corresponding β-coefficients for the four least-square linear regressions are shown.

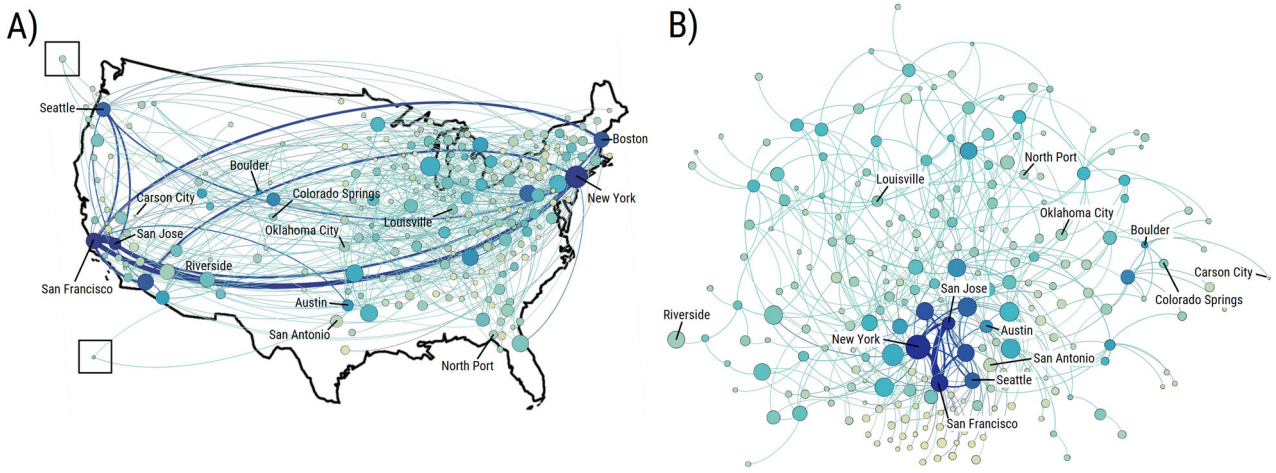

Fig. 4 The backbone of the Workforce Mobility Network (WMN) visualized. A On the map of the United States; and **B** on a standard force-directed network layout. A node's size is proportional to the area's population size, its color intensity is proportional to its PageRank centrality, and an edge's width is proportional to its weight.

so-called *β-coefficient* with each independent variable such as the sum of all independent variables multiplied by their respective β-coefficients approximates the value of the dependent variable with minimal error. Specifically, we used an ordinary least-square (OLS) regression model to estimate the coefficients such that the sum of the squared residuals between the estimation and the actual value is minimized. In line with what discussed by Bettencourt et al. (2010), it is more appropriate to express the dependent variable using absolute values

Table 1 Cumulative acquisition prices $\mathcal{A}_i$ for the ten metropolitan areas (MSAs) with the highest values of centrality/population ratio η, and the ten cities with the lowest values.

Top 10 according to η				Bottom 10 according to η			
MSA	**Population**	**Price $\mathcal{A}_i$**	η	**MSA**	**Population**	**Price $\mathcal{A}_i$**	η
San Jose, CA	1.8M	245B	11.4	Riverside, CA	4.2M	315M	0.044
San Francisco, CA	4.3M	160B	6.69	Columbia, SC	0.8M	2M	0.124
Boulder, CO	0.3M	1.7B	6.06	Oklahoma, City, OK	1.2M	0.6B	0.148
Boston, MA	4.5M	159B	3.00	Cape Coral, FL	0.6M	3.6M	0.153
Ithaca, NY	0.1M	0.5B	2.56	Fresno, CA	1.0M	29M	0.163
Austin, TX	1.7M	14B	2.43	Chattanooga, TN	0.5M	160M	0.174
Durham, NC	0.5M	3.4B	2.42	Virginia Beach, VA	1.7M	4.7B	0.177
Ann Arbor, MI	0.3M	0.4B	2.12	Buffalo, NY	1.1M	27M	0.185
Seattle, WA	3.4M	32B	2.09	Greensboro, NC	1.1M	0.6B	0.187
Trenton, NJ	0.3M	14B	1.94	York-Hanover, PA	0.4M	0.4B	0.201

(i.e., number of successful startups, total acquisition prices) rather than using ratios (e.g., percentage of successful startups) or per-capita values. That is because these two latter quantities implicitly assume that the dependent variable (e.g., innovation measure) linearly increases with the independent variables (e.g., number of existing startups, population size), while we know that it tends to super-linearly increase with them. Since all regression variables had skewed distributions, we log-transformed them using base-10 logarithm.

In the regression models, we experimented with two different groups of predictors (whose cross-correlations are shown in *Supplementary Information*): (i) socio-economic indicators; and (ii) indicators based on WMN's structure. First, the socio-economic indicators based on the literature are population size (Bettencourt et al., 2007a), population density (Jacobs, 1961), and number of patents granted in each metropolitan area (Bettencourt et al., 2007) in the year of 2010. To those three indicators, we added two others derived from CrunchBase: the number of active startups $\mathcal{N}_i$ in 2010, and the total past funding $\mathcal{F}_i$ raised up to the year of 2010. The number of active startups $\mathcal{N}_i$ is an upper bound for the number of successful ones and, as such, represents an important variable to control for; on the other hand, the independent variable of past funding $\mathcal{F}_i$ is not necessarily correlated with our dependent variable (i.e., with the actual innovation levels of companies), can be influenced by factors such as local tax policies, and, as such, can be regarded as a proxy for innovation incentives each area tends to enjoy.

Second, the indicators based on WMN's structure aim at capturing each area's centrality in the flows of ideas, techniques, knowledge, creative inputs, and business opportunities (Moreno et al., 2020). To characterize the potential exposure of a metropolitan area to these flows, we computed four centrality measures: degree centrality, node strength, Google PageRank, and harmonic closeness (see *Methods*). If we imagine knowledge as a collection of discrete units and assume that these units randomly flow in WMN, then an area's PageRank score is the fraction of the global knowledge the area has potential access to (e.g., if the score is 0.2, then 20% of the global knowledge is potentially accessible by the area). In a similar way, area *i*'s harmonic closeness is the distance (measured as the weighted number of hops) that a given unit of information needs to traverse to reach node *i* starting from any other node (Boldi and Vigna, 2014; Crucitti et al., 2006; Marchiori and Latora, 2000; Pan and Saramäki, 2011).

Table 2 reports the adjusted coefficients of determination R^2 and the β-coefficients for the ten models. The first 9 models consider the independent variables separately. We see that predicting acquisition prices $\mathcal{A}_i$ is harder than predicting the number of successful startups $\mathcal{S}_i$, yet the relative power of the predictors is mostly consistent across the two innovation measures. All the socio-economic indicators (models 1–5) are good predictors for the two measures, and, among them, the control variable of the number of active startups (5) is the most powerful predictor for the number of successful startups $\mathcal{S}_i$ ($R^2 = 0.92$) and is among the most predictive variables for the cumulative acquisition prices $\mathcal{A}_i$ ($R^2 = 0.57$). That is also because the number of active startups is an upper bound for the number of successful ones. In line with previous empirical findings (Bettencourt et al., 2007a), population (1) is positively correlated with both innovation measures. However, population density (2) is less so. Past fundings (3) and number of patents (4) are also positively associated, yet have the smallest β-coefficients. The last four models (models 6–9) test our four network centrality measures: PageRank (6) and node strength (7) have higher β-coefficients and R^2 compared to node degree, which do not account for network weights (8), and harmonic centrality (9). Overall, PageRank outperforms population size by 23% when predicting the number of successful startups $\mathcal{S}_i$, and is the top predictor of the cumulative acquisition prices $\mathcal{A}_i$, outperforming population by 36%.

To further disentangle the unique contribution of each predictor, we used a stepwise feature selection procedure to select the combination of predictors with the highest R^2. Specifically, we used the *stepAIC* algorithm implemented in the R standard packages, a widely used search method for feature selection. The method is based on the Akaike Information Criterion (Sakamoto et al., 1986) (AIC), an estimate of the relative amount of information lost by a model to represent the process that generated the empirical data. The AIC score rewards models that achieve a high goodness-of-fit score and penalizes them if they become overly complex. *stepAIC* measures the AIC score of models obtained by removing different sets of features from the original model and selects the feature combination that yields the lowest AIC. The two models that consist of the selected variables are reported in column 10 in Table 2. PageRank is the only network metric retained by the feature selection method because it is the only one that, in combination with the socio-economic features, improves the overall prediction. Also, the β-coefficient of PageRank is the highest for $\mathcal{A}_i$, and the second highest (only after the control variable of the number of active startups) for $\mathcal{S}_i$. In both cases, the coefficients of determination are significantly larger than those obtained for the other variables, especially than those obtained for population size and density. The variability explained by these models is equal to that explained by either of the two models (columns "all" in Table 2) whose predictors consist of all the variables under study.

To then check whether these effects are not due to chance, we generated a null configuration by randomizing the values of each of the innovation metrics $\mathcal{A}_i$ and $\mathcal{S}_i$, and applied the best performing regression model to this null configuration (column "random" in Table 2). The result is that R^2 drops to zero, and all the coefficients are not statistically significant.

Table 2 β-coefficients of the regression models to predict the two dependent variables of innovation performance, namely the number of successful startups $\mathcal{S}_i$ and the cumulative acquisition price $\mathcal{A}_i$.

	Dependent variable: number of successful startups $\mathcal{S}_i$											
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(Random)**	**(All)**
Population size	1.247***									0.160**	0.450	0.127*
	(0.067)									(0.074)	(0.310)	(0.076)
Population density		1.208***								0.072	−0.203	0.064
		(0.131)								(0.054)	(0.225)	(0.053)
Past funding			0.580***							0.087***	0.040	0.097***
			(0.026)							(0.032)	(0.134)	0.033
Patents				0.851***								0.011
				(0.044)								(0.047)
Active startups					1.023***					0.423***	−0.304	0.462***
					(0.026)					(0.108)	(0.450)	(0.113)
Network PageRank						1.083***				0.392***	−0.045	0.510***
						(0.031)				(0.089)	(0.369)	(0.181)
Network strength							0.785***					0.000
							(0.025)					(0.307)
Network degree								1.749***				0.253
								(0.076)				(0.256)
Harmonic centrality									1.345***			−0.407
									(0.055)			(0.220)
Constant	−6.153	5.910	−3.662	−1.137	−0.869	4.186	−0.783	2.538	−3.536	0.032	−2.067	1.950
	(0.395)	(0.511)	(0.220)	(0.125)	(0.055)	(0.087)	(0.068)	(0.063)	(0.631)	(2.627)	(0.087)	(1.203)
Adjusted R^2	0.73	0.39	0.79	0.74	0.92	0.90	0.88	0.80	0.82	0.94	0.01	0.94

	Dependent variable: cumulative acquisitions prices $\mathcal{A}_i$											
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(Random)**	**(All)**
Population size	1.592***									0.268	−0.423	0.161
	(0.155)									(0.222)	(0.354)	(0.326)
Population density		1.478***										−0.114
		(0.242)										(0.229)
Past funding			0.743***									0.064
			(0.067)									(0.142)
Patents				1.151***						0.226	0.086	0.320
				(0.098)						(0.183)	(0.291)	(0.204)
Active startups					1.316***							0.211
					(0.099)							(0.487)
Network PageRank						1.442***				1.05***	0.065	1.815**
						(0.103)				(0.225)	(0.359)	(0.782)
Network strength							1.017***					−0.498
							(0.079)					(1.321)
Network degree								2.243***				0.814
								(0.198)				(1.105)
Harmonic centrality									1.710***			−1.150
									(0.150)			(0.947)
Constant	−0.412	14.70	2.753	5.808	6.310	12.94	6.402	10.69	2.952	9.661	11.43	16.67
	(0.919)	(0.945)	(0.563)	(0.280)	(0.211)	(0.288)	(0.212)	(0.165)	(0.535)	(1.765)	(2.816)	(5.182)
Adjusted R^2	0.44	0.22	0.48	0.51	0.57	0.60	0.55	0.49	0.49	0.60	0.00	0.61

Standard errors for the coefficients are reported in parenthesis. The *random* column refers to the best performing regression model applied to a null configuration where the values for each innovation metric ($\mathcal{S}_i$ for the top table, and $\mathcal{A}_i$ for the bottom one) are randomized across areas.
*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

In multivariate regressions, if the independent variables are perfectly independent, then the coefficient of determination R^2 decomposes itself into the sum of the squares of the Pearson's correlation coefficients computed for each variable separately. However, in our case, as in the majority or real-world scenarios, most of the variables are correlated with each other, and the sum of each independent R^2 exceeds the one obtained for the multivariate regression (model 10). To properly decompose the relative contribution of the correlated independent variables, we used the Lindeman, Merenda and Gold (LMG) method (Lindeman et al., 1980) and computed the relative importance of each predictor (Fig. 5). To estimate the feature importance, we used the implementation of the LMG method provided in R in the package *relaimpo* (Grömping, 2006). LMG estimates the proportion of the R^2 contributed by each individual predictor by adding the predictors to the regression model sequentially. The increased R^2 represents the contribution by the predictor added. Since the sequence of feature addition influences the R^2 increase, LMG averages the value of the contributions across all possible feature orderings. Interestingly, after controlling for the number of active startups, PageRank is confirmed to be the predictor that explains most of the variability in the data.

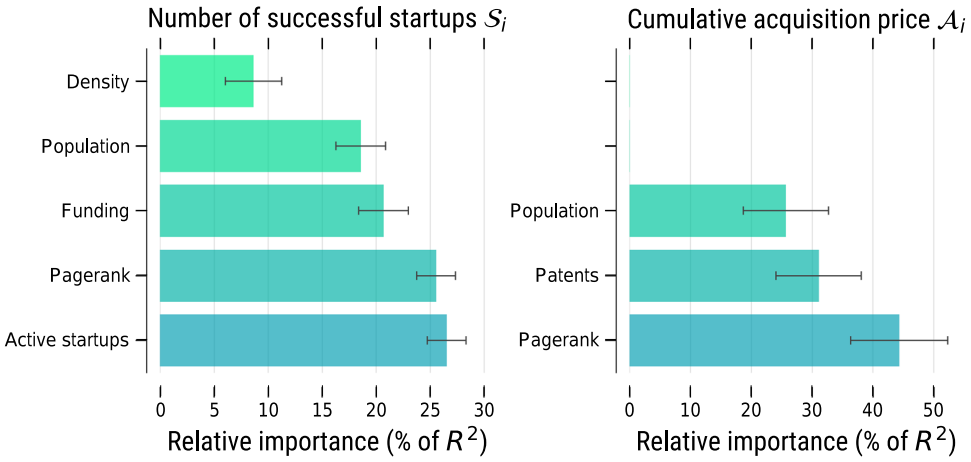

Fig. 5 Relative contribution of each variable to the total variance explained by the regression models in Table R. The relative contribution is expressed as percentage of the R^2 determined by each variable.

Discussion

To place our results in a broader context, consider that we have corroborated previous work in that we have found similar superlinear scaling relations between our innovation metrics and city size (Arbesman et al., 2009; Arcaute et al., 2015; Bettencourt et al., 2007a, b). Such work has typically attributed superlinear scaling relations to mainly one *endogenous* factor: that of increased social interconnectivity within cities (an emergent property of city life). This is the most widely accepted explanation in the literature. Yet the very same work has also conceded that there are other *exogenous* factors that could further explain higher levels of innovation in cities. Indeed, with city size, there have been observed significant changes in, for example, the ability of disproportionately attract talent (Florida, 2005; Glaeser, 2011; Keuschnigg et al., 2019).

Our findings complement the widely accepted explanation of "increased social interconnectivity in cities" by offering a more nuanced understanding of urban innovation. We find that our metrics of workforce mobility, albeit imperfect, predict innovation levels that were previously unexplained by superlinear growth. Despite what a scaling relationship suggests, a percentage increase in population size might not be necessarily followed by a percentage change in innovation. That is because big cities do not grow in random ways but grow through their *selective* attraction of talent(Keuschnigg et al., 2019). On a policy level, this should bring a fundamental shift of focus: from blind city growth to *selective* city growth. Ideally, policies should enable selective processes that are considered desirable (e.g., those resulting in the attraction of talent without suffering from the consequences of urban displacement and gentrification). Economists have put forward quantitative evidence, suggesting that a city's economic performance is also influenced by the type of people who migrate to the city (e.g., by the migration of the so-called "creative class" (Florida, 2005)), and they have typically done so based on migration records (Keuschnigg et al., 2019); yet, these records do not differentiate the variety of migration flows, let alone the types of workforce flows that support the emergence of *new entrepreneurial ecosystems*.

Based on these previous findings, we hypothesized that the network of informal interactions between professional working at startups who carry their expertize as they move from one city to another is predictive of innovation outcomes. This is the first study that has built a Workforce Mobility Network at the scale of an entire country from open data, and that has shown that this network's structural characteristics are predictive of urban innovation: global network measures tend to predict long-term innovation better than even what cumulative investments do.

Our study comes with limitations that are mostly determined by our data. No sufficient longitudinal data was available for testing causal relationships and for ascertaining the robustness of the model across historical periods characterized by different patterns of economic activity. Furthermore, startups do not have to publicly disclose their funding rounds or acquisition prices: 83% of the funding rounds in our dataset, for example, have been fully disclosed on CrunchBase. Yet, as shown in *Supplementary Information*, being of random nature, such missing data has little impact on our two innovation measures, and no impact on a comparative evaluation of areas. Finally, the time frames over which workforce mobility and urban innovation were measured did not necessarily overlap. As one expects, the more up-to-date the workforce mobility data, the higher its predictive power. Yet, as reported in *Supplementary Information*, our two urban innovation measures could still be accurately predicted from workforce mobility data that was 5 years older. When using workforce mobility data up to 2005 only, we could predict the number of successful startups S_i and the cumulative acquisition price A_i with an adjusted R^2 of 0.56 and one of 0.67, respectively —compared to 0.60 and 0.75 obtained by using the data up to 2010.

Data availability

All the datasets used in this work can be fully and freely downloaded from the Web. The CrunchBase data is available through its public API at https://data.crunchbase.com, patent data can be downloaded from http://www.patentsview.org/download, and US census data from https://www.census.gov. To map CrunchBase firms to metropolitan areas, we used the census data available here: https://www.census.gov/geographies/reference-files/time-series/geo/relationship-files.html. An interactive visualization of the network data is available on the project's website at http://goodcitylife.org/cities4innovation.

Received: 1 July 2020; Accepted: 25 November 2020;
Published online: 07 January 2021

Note

1 An interactive visualization of the network is available on the project's website at http://goodcitylife.org/cities4innovation.

References

Acs ZJ, Mueller P (2008) Employment effects of business dynamics: mice, gazelles and elephants. Small Bus Econ 30:85–100

Arasu A, Novak J, Tomkins A, Tomlin J (2002) Pagerank computation and the structure of the web: experiments and algorithms. In Proceedings of the Eleventh International World Wide Web Conference, Poster Track, ACM, pp. 107–117

Arbesman S, Kleinberg JM, Strogatz SH (2009) Superlinear scaling for innovation in cities. Phys Rev E 79:016115

Arcaute E et al. (2015) Constructing cities, deconstructing scaling laws. J R Soc Interface 12:20140745

Barbosa H et al. (2018) Human mobility: models and applications. Phys Rep 734:1–74

Barthelemy, M. The structure and dynamics of cities. Cambridge University Press, 2016.

Barthelemy M (2019) The statistical physics of cities. Nat Rev Phys 1:406–415

Bettencourt LM, Lobo J, Helbing D, Kühnert C, West GB (2007a) Growth, innovation, scaling, and the pace of life in cities. Proc Natl Acad Sci USA 104:7301–7306

Bettencourt LM, Lobo J, Strumsky D (2007b) Invention in the city: increasing returns to patenting as a scaling function of metropolitan size. Res Policy 36:107–120

Bettencourt LM, Lobo J, Strumsky D, West GB (2010) Urban scaling and its deviations: revealing the structure of wealth, innovation and crime across cities. PLoS ONE 5:e13541

Boldi P, Vigna S (2014) Axioms for centrality. Internet Math 10:222–262

Bos JW, Stam E (2014) Gazelles and industry growth: a study of young high-growth firms in the netherlands. Ind Corp Chang 23:145–169

Burt RS (1993) The social structure of competition. Explor Econ Sociol 65:103

Cohen L, Gurun UG, Kominers SD (2016) The growing problem of patent trolling. Science 352:521–522

Coscia M, Neffke FM (2017) Network backboning with noisy data. In 2017 IEEE 33rd International Conference on Data Engineering (ICDE), IEEE, pp. 425–436

Crucitti P, Latora V, Porta S (2006) Centrality measures in spatial networks of urban streets. Phys Rev E 73:036125

Decker R, Haltiwanger J, Jarmin R, Miranda J (2014) The role of entrepreneurship in us job creation and economic dynamism. J Econ Perspect 28:3–24

Depersin J, Barthelemy M (2018) From global scaling to the dynamics of individual cities. Proc Natl Acad Sci 115:2317–2322

Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328:1029–1031

Florida, R. Cities and the creative class. Routledge, 2005

Glaeser E (2011) Triumph of the city: how urban spaces make us human. Pan Macmillan

Glaeser E, Scheinkman, J (2001) Measuring social interactions. In: Durlauf, SN and Young, HP (eds) Social dynamics, ch. 4. MIT Press, Boston, MA. pp. 83–132

Glaeser EL, Rosenthal SS, Strange WC (2010) Urban economics and entrepreneurship. J Urban Econ 67:114

Grömping U et al. (2006) Relative importance for linear regression in r: the package relaimpo. J Stat Softw 17:1–27

Hall PG, Raumplaner S (1998) Cities in civilization. Pantheon Books, New York

Haltiwanger J, Jarmin RS, Miranda J (2013) Who creates jobs? small versus large versus young. Rev Econ Stat 95:347–361

Hargadon AB (1998) Firms as knowledge brokers: lessons in pursuing continuous innovation. California Manag Rev 40:209–227

Jacobs, J (1961) The death and life of great American cities. Vintage

Jacobs, J (1970) The economy of cities. economics & sociology. Vintage Books

Keuschnigg M, Mutgan S, Hedström P (2019) Urban scaling and the regional divide. Sci Adv 5:eaav0042

Kirkley A, Barbosa H, Barthelemy M, Ghoshal G (2018) From the betweenness centrality in street networks to structural invariants in random planar graphs. Nat Commun 9:1–12

Lämmer S, Gehlsen B, Helbing D (2006) Scaling laws in the spatial structure of urban road networks. Phys A 363:89–95

Latora V, Nicosia V, Russo G (2017) Complex networks: principles, methods and applications. Cambridge University Press.

Lindeman R, Merenda P, Gold R (1980) Introduction to bivariate and multivariate analysis. Scott, Foresman, & Co, New York

Makarem NP (2016) Social networks and regional economic development: the los angeles and bay area metropolitan regions, 1980–2010. Environ Plan 34:91–112

Marchiori M, Latora V (2000) Harmony in the small-world. Phys A 285:539–546

Moreno B et al. (2020) Predicting success in the worldwide start-up network. Sci Rep 10(1): 345

Moretti E (2012) The new geography of jobs. Houghton Mifflin Harcourt

Mumford L (1961) The city in history: its origins, its transformations, and its prospects, vol 67. Houghton Mifflin Harcourt

Nicholas T (2013) Are patents creative of destructive. Antitrust LJ 79:405

Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web, Stanford

Pan RK, Saramäki J (2011) Path lengths, correlations, and centrality in temporal networks. Phys Rev E 84:016105

Parise S, Whelan E, Todd S (2015) How twitter users can generate better ideas. MIT Sloan Manag Rev 56:21

Powell WW, Koput KW, Smith-Doerr, L (1996) Interorganizational collaboration and the locus of innovation: networks of learning in biotechnology. Administ Sci Quart 41(1):116–145

Sakamoto Y, Ishiguro M, Kitagawa G (1986) Akaike information criterion statistics. vol. 81. D. Reidel, Dordrecht, The Netherlands

Saxenian A, A.C. of Learned Societies (1996) Regional advantage: culture and competition in silicon valley and route 128, with a new preface by the author. Harvard University Press

Sorenson O, Stuart TE (2001) Syndication networks and the spatial distribution of venture capital investments1. Am J Sociol 106:1546–1588

Tria F, Loreto V, Servedio VDP, Strogatz SH (2014) The dynamics of correlated novelties. Sci Rep 4:1–8

Weins J, Jackson C (2014) The importance of young firms for economic growth. Entrepreneurship Policy Digest

Xing W, Ghorbani A (2004) Weighted pagerank algorithm. In: Proceedings of Second Annual Conference on Communication Networks and Services Research, IEEE, pp. 305–314

Acknowledgements

We thank Valerio Ciotti for his help in collecting the data. VL work was funded by the Leverhulme Trust Research Fellowship "CREATE: the network components of creativity and success".

Author contributions

MB collected the data. MB and LMA conducted the experiments and analyzed the results. All authors conceived the experiments and contributed to write the manuscript.

Competing interests

The author(s) declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1057/s41599-020-00685-7.

Correspondence and requests for materials should be addressed to D.Q.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.