In *Advances in Network Science*, A. Wierzbicki, U. Brandes, F. Schweitzer, edts,
Proc. Int. Conf. NetSci-X 2016, Wroclaw, Poland, Jan 11-13, 2016,
Lecture Notes in Computer Science, vol. **5964**, 2016, Berlin, pp. 178-185.

Social Ties as Predictors of Economic Development*

Buster O. Holzbauer[1], Boleslaw K. Szymanski[1], Tommy Nguyen[1], Alex Pentland[2]

[1] Social Cognitive Network Academic Research Center (SCNARC),
Rensselaer Polytechnic Institute, Troy NY 12180, USA,
holzbh@rpi.edu
[2] Massachusetts Institute of Technology, Cambridge MA 02139, USA

Abstract. A social network is not only a system of connections or relationships, but pathways along which ideas from various communities may flow. Here we show that the economic development of U.S. states may be predicted by using quantitative measures of their social tie network structure derived from location-based social media. We find that long ties, defined here as ties between people in different states, are strongly correlated with economic development in the US states from 2009-2012 in terms of GDP, patents, and number of startups. In contrast, within-state ties are much less predictive of economic development. Our results suggest that such long ties support innovation by enabling more effective idea flow.

1 Introduction and Related Work

Studies in economic sociology suggest that peer-to-peer human relationships affect economic opportunities because information about these opportunities often spread most effectively between people [7, 9, 10, 15, 16, 23, 24]. Information spreading via interpersonal relationships is often richer than traditional broadcast media such as television, newspaper, radio, etc. because acquaintances can interact face-to-face, provide relevant information when needed, and influence one another with respect to adopting new behavior and ideas [22].

It has been argued that information coming from weak ties is often richer than information arriving via strong ties because "those to whom we are weakly tied are more likely to move in circles different from our own ... and have access to information different from what we [usually] receive [16]." Weak ties have been shown to be valuable sources of information because individuals can use them

* Research was partially sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA) and by the Office of Naval Research Contract N00014-15-1-2641. The content of this paper does not necessarily reflect the position or policy of the U.S. Government, ARL or ONR, no official endorsement should be inferred or implied.

to find jobs [7, 15], solicit feedback on starting new ventures [24], and search for people like in the small-world experiment [4, 17, 18, 25]. In other settings such as examining workplaces, social network structure can affect productivity and innovation of employees and could lead to higher compensation, more promotion opportunities, and better performance evaluations [9, 10, 23, 24]. Therefore, the effect of weak ties on economic opportunities suggests that perhaps the number and distribution of social ties might also be used for measuring economic development on a larger scale.

Contemporary research on urban characteristics and growth has demonstrated scaling laws for innovation and wealth creation as a power function of the population size as expressed by the equation: $y(t) = cx(t)^m$ where $x(t)$ is the population size and $y(t)$ is the metric of innovation at time t [5, 6]. These results show that as the population size increases, GDP, wages, patents, private research employment and development increase at superlinear rates where $1.03 \leq m \leq 1.46$ [6]. Perhaps the best explanation for the superlinear scaling of wealth creation is that as the population size increases, the density of social relationships between people increases because there are more choices for establishing relationships [21]; therefore, increasing the connectivity between people decreases the time for ideas to spread.

Following this line of thinking, recent results in [21] suggest that a generative model for tie formation as a function of social tie density yields somewhat better results than purely descriptive models based only on population size, and in addition offers a simple causal theory of these scaling phenomena. Results obtained under modest assumptions (nodes distributed uniformly on a Euclidean space, connections established following the rank friendship model [18]) show that algorithmically generated social ties based on social tie density can be used to model urban characteristics of cities such as GDP, number of patents, research employment, etc.

Here we extend this line of thinking by focusing on characteristics of economic development as a function of idea flow based on peer-to-peer social relationships and find that "long ties" (defined below) are a main component enabling such flow. This was accomplished by using data containing geographical locations and friendship information of hundreds of thousands of people from location-based social media, namely Gowalla [20]. Also, these datasets allow us to infer face-to-face interactions [19] and measure the strength of ties in terms of not only interactions but also geographical and "administrative" distances (i.e., short or long ties [11, 14]).

Other approaches for measuring economic development of large geographical areas include examining the diversity of social contacts (i.e., call detail records as a proxy for social relationships) since more contacts imply more channels for receiving information [13]. Yet using calling patterns to infer social contacts is biased towards those that are more likely to be connected via strong ties since weak ties are by definition those that are used infrequently. While these approaches [13, 21] can vary in their methodologies, ranging from mathematically oriented to data-driven, what they share in common is using social network

analysis to predict innovation, wealth creation, and other patterns of complex human behavior. In this paper, the novelty of our approach lies at the intersection of economic sociology (i.e., the interplay of long ties and economic opportunities) and simple contagion models (i.e., the spread of ideas from one place to another). Results show that the speed of access to ideas is a strongly correlated with social diversity and also a signature of the economic development of US states without needing to tune parameters or incorporate secondary factors such as the level of educational attainment and internal transportation infrastructure.

2 Data

Our primary focus in this paper is the Gowalla dataset detailed in previous publication [20]. The reasoning behind using Gowalla is that a location-based social network allowed us to analyze both social interactions and geographic interactions separately (through friendships and check-ins respectively). In this paper we considered specifically applying the network towards modeling U.S. GDP [1], patents [3], and small startups (20 or less employees) [2], so we removed any users and corresponding friendship links that were not internal to the United States. This left us with 75,803 users, 464,556 "long ties" (defined as friendships where the two users were in different physical U.S. states), and 222,072 "short ties" (friendships where users were in the same state).

In this paper we do not discuss our model for idea flow, but note that by examining correlations between GDP, patents, startups, and idea flow, we found a near-perfect match between correlations using long ties and simulated idea flow. For this reason, we elected to do the rest of our analysis and discussion here using long ties as a proxy for idea flow. This is advantageous since long ties can be observed directly from the network structure, so there is less uncertainty in the accuracy of analysis based on long ties. For a given state i, we define its census population as P_i, the number of long ties L_i as the number of ties with one end in another state, and the number of short ties (edges) entirely within the state as S_i.

In addition we also considered a community-detection (network clustering) approach, however due to the space limitations and their much lower correlations, we chose to exclude the results based on "bridges" between communities from this paper. The correlations we found for community bridges were very similar to those of the short ties discussed here, though the reasons that both bridges and short ties poorly match our economic metrics of interest may be unrelated. In contrast, idea flow could be formally calculated based on long ties, and this is why we are comfortable claiming that long ties can be used as a simple and accurate substitute for more direct but difficult methods of representing flow of ideas.

3 Methods and Results

The first thing we examined was how indicators P_i, L_i, and S_i correlate with metrics GDP_i, $Patents_i$, and $Startups_i$. The results shown in Table 1 indicate that population is better correlated with the metrics then either type of ties, and short ties correlations are particularly low.

Table 1. Correlations Between Indicators and Economic Metrics

Feature	GDP	Patents	Startups
Population	.985	.865	.982
Long Ties	.921	.788	.892
Short Ties	.692	.531	.599

Such high correlations of total population can arise because either each additional person adds a similar increment to the network of social relationships and idea flow, or their individual cognitive processes are generating innovations independent of their social context. Thus, it is interesting that short ties (within the same state) are relatively less correlated with the metrics, while long ties (between states) have correlations that are significantly stronger.

We therefore examined P_i, L_i, and S_i in the context of distributions over each indicator and computed the probability that state data are drawn from them. Moreover, we looked how this probability changes as we enrich models by adding successively more indicators. For the sake of space we omit here details of the models, but based on a linear model we estimated Gaussian distributions for each economic metric against single variables (P,L,S models), pairs (PL,PS,LS models), and a three-variable model (PLS) using Maximum Likelihood Estimation [12]. This estimation was computed by approximating the likelihood function derivative solution to zero and then following the highest gradient descent to the nearest maximum, so we cannot guarantee that we found the global extrema. The logs of maximum likelihoods of fitting state data by each model are shown in Table 2.

From examining the likelihood ratios, we can find the probability that the two models are not the same via the Likelihood Ratio Test (LRT) [12]. This method works when the compared models are nested (one model's parameters are a subset of the other model's parameters). In this case, a Chi-Square distribution with the degree of freedom equal to the difference in the number of parameters between the models can be used to find confidence level with which we can conclude if the models are different. For cases where the models are not nested, we instead apply the Akaike information criterion (AIC) [8], in which case we require a difference in AIC of around 3.0-4.0 (depending on the number of parameters), which can be derived from standard log-normal distribution tables. The AIC of the model is defined as $-ln(L)+2(p+1)$ where L is the likeli-

Table 2. MLE Of Indicators Fit To Economic Metrics

Feature	GDP	Patents	Startups
S	-691.53	-451.31	-667.43
L	-665.86	-435.15	-641.84
LS	-660.96	-434.82	-641.02
P	-632.15	-425.11	-576.98
PS	-632.15	-417.08	-575.24
PL	-609.46	-425.16	-576.62
PLS	-604.33	-417.08	-575.24

Table 3. MLE Differences For Confidence Levels Using LRT

Confidence Level:	0.95	0.99	0.999
ΔDegrees of Freedom = 1:	1.92	3.32	5.5
ΔDegrees of Freedom = 2:	3.00	4.61	6.9

hood of fitting the state data with the model and p is its number of parameters, as shown in Table 3.

Using this methodology, we find that the joint PL model noticeably benefits from information provided by long ties for GDP. The improvement is so significant that it is likely to result from information contributed by the long ties and not captured by the population alone. In contrast, the difference of likelihoods between PS model which includes short ties and population-only P model is not statistically significant for GDP and Startups. The same is true for the LS and L models for Patents and Startups, and since L is statistically significantly better than S model, this means that long ties alone capture all features that make LS superior to the S model. Moreover in all cases of independent variables, long ties alone are significantly better than short ties for all three independent variables. Because of the nearly exact match of long ties and our simulation of idea flow, the same should be true of other measurements of idea flow. As a summary, the list of models for which the differences in likelihoods are not statistically significant is: P and PS models for GDP, L and LS for Patents and Startups, P and PL as well as PS and PLS for Patents and finally P, PS, PL, and PLS for Startups.

4 Discussion

From our observations, it appears that that productivity and innovation at the state level within the US are more about connecting different states than bridging across different local communities operating in the same state. When taken together with the fact that idea flow accounts for the super-linear scaling of cities, and that long ties are nearly perfectly correlated with simulations of idea

flow across the entire US, these results support the hypothesis that idea flow between states is a major source of state level innovation and productivity.

This conjecture that it is idea flow between separated communities that accounts for state-level economic variations is supported by the significant increase in model matches to data for both GDP and patents that are obtained when network structure is added to population information. The fact that adding long ties to population model increases the probability of the extended model fitness suggests that the correlation between long ties and the economic metrics is due to a different phenomenon than that associated with simple variation in population.

In the Granovetter paper cited earlier, the authors discussed that there are many criteria one can use to define strength of a tie. Even within community detection there are many decisions to be made, for example depending on the algorithm, there may be room for overlapping communities, thresholds that can be changed, or different methods of weighting ties. We find it telling that there is a disparity in the fraction of long ties that are weak and the fraction of short ties that are weak, and believe that this explains why long ties improved our economic predictions more than short ties. It is a particularly attractive idea since it encodes idea flow across borders, which a social network could contribute, but raw population values would not.

It is also important to note that our subject sample were users of Gowalla, both because users of Gowalla must have more than average disposable income in order to be able to possess a smartphone and be innovative enough to embrace technology that was at that time quite new and make use of such a location-based social network. We believe that economic performance such as GDP or having startups is furthered by the advancement and utilization of technology, and so the Gowalla userbase may be a more appropriate sample than the U.S. population as a whole. We do not, of course, believe that the Gowalla population is a representative sample of the entire population, but rather a sample that is well suited to predicting the economic factors we examined.

Communities were still useful as one way to measure strength of ties. While long ties and short ties are not directly analogous to the concept of strong and weak ties, our thought process was that long and short ties might some of the same properties as strong and weak ties. To test this intuition, we ran community detection using GANXiS [26], and defined a pair of users as having a strong tie if they were in the same community, and otherwise we considered the pair to have a weak tie. We summarize information about ties in the Gowalla component that we use in Table 4; clearly, nearly the same fractions of short and long ties are weak and close to the fraction of weak ties among all ties. Since, as we show later, long ties perform better than short ones, we expect that long ties will also outperform weak ties as predictors of economic metrics.

Table 4. Summary of Geographic Ties and Strength-Based Ties

Total Short	444144	Total Short and Weak	308132
Total Long	929112	Total Long and Weak	723900
% of Short Ties that are Weak	69.38	% of Ties that are Weak	75.15
% of Long Ties that are Weak	77.91	% of Ties that are Long	67.86

5 Conclusion

GDP, patents, and startups are three economic measurements that can be used to quantify productivity and innovation. By modeling these measurements using location-based social network data, we find that not only do we get a linear relationship with high correlation, but also that the long tie network produces this correlation through different means than the population-only model. While correlation is not causation, there is intuition to support a conjecture that the long tie network features are connections that allow diverse ideas to be shared among individuals. Since ideas may be readily shared among individuals in a particular geographic region due to shared culture and higher probability of regular interaction, long ties are an especially good candidate for measuring the speed of sharing of novel ideas because they connect people acting in separate innovation support infrastructures of different states.

Our results indicate that while we see improvements by combining long ties and population for GDP and patent prediction, we do not see the same behavior for predicting startups. One plausible explanation why startups behave differently is that only a small percentage of startups are innovation-based, while the majority are self-employed individuals providing standard personal services. We plan to verify this hypothesis in future work. In the future we also intend to expand on the other probability distributions we looked at, additional network features and measurements derived from network features, and provide the rigorous mathematical derivations that lead to our parameter estimation and MLE bounding.

References

1. U.S. bureau of economic analysis. `http://www.bea.gov/newsreleases/regional/gdp_state/2015/xls/gsp0615.xlsx`, accessed: 2015-09-28
2. U.S. census bureau, statistics of u.s. businesses (susb). `http://www2.census.gov/econ/susb/data/2012/us_state_totals_2012.xls`, accessed: 2015-09-28
3. U.S. patent and trademark office, patent technology monitoring team. `http://www.uspto.gov/web/offices/ac/ido/oeip/taf/cst_utl.htm`, accessed: 2015-09-28
4. Adamic, L., Adar, E.: How to search a social network. Social networks 27(3), 187–203 (2005)
5. Bettencourt, L., West, G.: A unified theory of urban living. Nature 467(7318), 912–913 (2010)

6. Bettencourt, L.M., Lobo, J., Helbing, D., Kühnert, C., West, G.B.: Growth, innovation, scaling, and the pace of life in cities. Proceedings of the national academy of sciences 104(17), 7301–7306 (2007)
7. Boxman, E.A., De Graaf, P.M., Flap, H.D.: The impact of social and human capital on the income attainment of dutch managers. Social networks 13(1), 51–73 (1991)
8. Burnham, K., Anderson, D.: Model selection and inference: a practical information-theoretic approach (1998)
9. Burt, R.S.: Structural holes and good ideas. American journal of sociology 110(2), 349–399 (2004)
10. Burt, R.S.: Structural holes: The social structure of competition. Harvard university press (2009)
11. Centola, D., Macy, M.: Complex contagions and the weakness of long ties1. American Journal of Sociology 113(3), 702–734 (2007)
12. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological) pp. 1–38 (1977)
13. Eagle, N., Macy, M., Claxton, R.: Network diversity and economic development. Science 328(5981), 1029–1031 (2010)
14. Ghasemiesfeh, G., Ebrahimi, R., Gao, J.: Complex contagion and the weakness of long ties in social networks: revisited. In: Proceedings of the fourteenth ACM conference on Electronic Commerce. pp. 507–524. ACM (2013)
15. Granovetter, M.: Getting a job: A study of contacts and careers. University of Chicago Press (1995)
16. Granovetter, M.: The impact of social structure on economic outcomes. Journal of economic perspectives pp. 33–50 (2005)
17. Kleinberg, J.: The small-world phenomenon: An algorithmic perspective. In: Proceedings of the thirty-second annual ACM symposium on Theory of computing. pp. 163–170. ACM (2000)
18. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A.: Geographic routing in social networks. Proceedings of the National Academy of Sciences of the United States of America 102(33), 11623–11628 (2005)
19. Nguyen, T., Chen, M., Szymanski, B.K.: Analyzing the proximity and interactions of friends in communities in gowalla. In: Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on. pp. 1036–1044. IEEE (2013)
20. Nguyen, T., Szymanski, B.K.: Using location-based social networks to validate human mobility and relationships models. In: Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on. pp. 1215–1221. IEEE (2012)
21. Pan, W., Ghoshal, G., Krumme, C., Cebrian, M., Pentland, A.: Urban characteristics attributable to density-driven tie formation. Nature communications 4 (2013)
22. Pentland, A.: Social physics: How good ideas spread-lessons from a new science (2014)
23. Reagans, R., Zuckerman, E.W.: Networks, diversity, and productivity: The social capital of corporate r&d teams. Organization science 12(4), 502–517 (2001)
24. Ruef, M.: Strong ties, weak ties and islands: structural and cultural predictors of organizational innovation. Industrial and Corporate Change 11(3), 427–449 (2002)
25. Watts, D.J., Dodds, P.S., Newman, M.E.: Identity and search in social networks. science 296(5571), 1302–1305 (2002)
26. Xie, J., Szymanski, B.: Labelrank: A stabilized label propagation algorithm for community detection in networks. In: Network Science Workshop (NSW), 2013 IEEE 2nd. pp. 138–143 (April 2013)