Community Computing: Comparisons between Rural and Urban Societies using Mobile Phone Data

Nathan Eagle
The Santa Fe Institute
1399 Hyde Park Rd
Santa Fe, NM 87501
Email: nathan@mit.edu

Yves-Alexandre de Montjoye
Universite Catholique de Louvain
Department of Mathematical Engineering
4 avenue Georges Lemaitre, B-1348 Louvain-la-Neuve, Belgium
Email: YvesAlexandre.de.Montjoye@gmail.com

Luís M. A. Bettencourt
The Santa Fe Institute
1399 Hyde Park Rd
Santa Fe, NM 87501
Email: lmbett@lanl.gov

***Abstract*—We present a comparative analysis of the behavioral dynamics of rural and urban societies using four years of mobile phone data from all 1.4M subscribers within a small country. We use information from communication logs and top up denominations to characterize attributes such as socioeconomic status and region. We show that rural and urban communities differ dramatically not only in terms of personal network topologies, but also in terms of inferred behavioral characteristics such as travel. We confirm the hypothesis for behavioral adaptation, demonstrating that individuals change their patterns of communication to increase the similarity with their new social environment. To our knowledge, this is the first comprehensive comparison between regional groups of this size.**

I. Introduction

By necessity, our understanding of the social interactions within a society is typically derived from sampling: we take detailed information about a subsection of the population and make our estimate based on the results we obtain from this relatively microscopic sample. To achieve this, most empirical studies in sociology, economics and other social sciences depend on surveys of the population [1], [2], which strive to eliminate biases, but remain more an art than a science. However, data about aggregate human interactions is increasingly becoming available, creating an opportunity to take a macroscopic approach to these efforts. Today, more than 4 billion people living in virtually every country on Earth are continuously generating massive amounts of data about their movements, relationships, and even financial transactions.

We present an analysis of the mobile call data records (CDR) for all 1.4M mobile phone subscribers within a small country over four years, between January 2005 and January 2008. For each call in the CDR, we have access to not only the time, duration, and recipient of the call, but also the location of the cellular towers associated with the call when it began and terminated. Additionally, we have coupled this data with regional census and airtime sharing data. This allows us to empirically test several previous hypotheses about the behavioral patterns that differentiate urban and rural communities, and the individuals who move between them.

This paper presents the first quantitative comparison of urban and rural communities based on a complete mobile phone graph of an entire nation. We begin with an overview of related work, detailing both the study of rural and urban societies within the sociology and social psychology literatures, as well as previous studies involving mobile phone data. We then provide a detailed description of the data and list both the individual and social network attributes associated with urban and rural communities. Individuals who have moved between urban and rural communities are identified and we measure how their different attributes change in response to their environment. Lastly, we conclude with a discussion on the potential of this type of data for a wide range of additional research questions.

II. Related Work

A. The Effects of Cities on Personal Networks

Many of the foundational concepts of sociology, social psychology and economics have originated from the observation that the transition of populations from rural areas to urban centers results in both behavioral and socioeconomic changes.

It has been observed long ago that the socioeconomic structure of societies changes as they urbanize [3], [4], [5], [7]. At one extreme, rural social relations are typically heavily predicated on kinship and on a logic of subsistence, where economies tend to be less diverse and individuals less interdependent. In the city, however social and economic interdependence becomes crucial, more utilitarian social relations develop, and greater economic opportunities are available, typically resulting on greater economic productivity and personal wealth, at least on the average, Simultaneously social psychologists [6] have emphasized the phenomenon of 'social overload' in relation to life in the larger cities. Social overload is the inability to respond or fulfill the many more potential social connections available to an individual than those he/she may have time to realize and explore.

These qualitative observations imply specific measurable consequences for the structure of social networks and to their geographic variation. Research in sociology [9], based on large survey data around San Francisco, has shown that personal networks change in systematic ways from small towns to the city and also that those changes are related to age, occupation and education. Cities reflect these changes, at least in part, because urban areas appear to attract younger and more educated people [8] who typically have larger and more

diverse personal networks [9]. Additionally, the strength of kin relations in some regions has been linked to socioeconomic status, while increased geographic distance does not necessary imply a diminished importance for providing social support [10].

B. Behavioral Studies using Mobile Phone Data

The recent analysis of data from mobile phone service providers have led researchers to increased insight into human movement patterns. While some researchers take issue with labeling these insights as 'universal laws of human movement', it is clear that through the analysis of cellular tower location data from hundreds of thousands of people, it is possible to finally quantify some of the more fundamental rules of human motion [11]. Other studies are also explicitly using mobile phone data to study the dynamics of cities in effort to better inform urban planning policies [16].

A different type of social data collected by mobile phone service providers is the adoption of service (tariff) plans and other telecommunication products, which may be thought of as the spread of a social contagion. In other studies, particular individuals are identified who hold influence over others in their peer group; when they adopt a product or service, the individuals whom they call subsequently adopt the product as well. Through the analysis of the diffusion of these social contagions over a call graph it may become possible to learn more about the social dynamics inherent within a population[14], [15].

III. Data and Methodology

A. Data Description

Mobile phone service providers have a wealth of movement and communication in their call data records (CDR). While obtaining access to these operator databases is not a trivial process for researchers, today's mobile phone service providers occasionally allow limited access to the anonymized data they log about their subscribers' behavior[13], [11], [12]. This data consists of all communication events (phone calls and text messages) as well as the cellular tower that enabled the communication to occur. It is important to emphasize the typical constraint on CDR: location of a phone is only logged if the phone is actively being used to communicate. This means that for the vast majority of times, the phone's location is unknown. While a mobile phone continuously monitors signals from proximate cellular towers, due to power constraints it typically does not continuously send back similar signals alerting the nearby towers of its particular location. [1]

The data used in this paper consists of four years of CDR for every mobile phone subscriber within a small country. However, as in previous research, we do not have access to phone numbers, but rather unique IDs that provide no personally identifiable information. Besides the standard information within CDR including voice and text-message communication and location estimates based on cellular towers, we also have access to additional subscriber data, including pre-paid scratch card denominations, air-time sharing, product adoption data, and phone model.

[1]Operators can also 'ping' a phone to have it report back to a nearby tower, however this requires additional power from the phone and therefore typically is impractical for continuous location tracking.

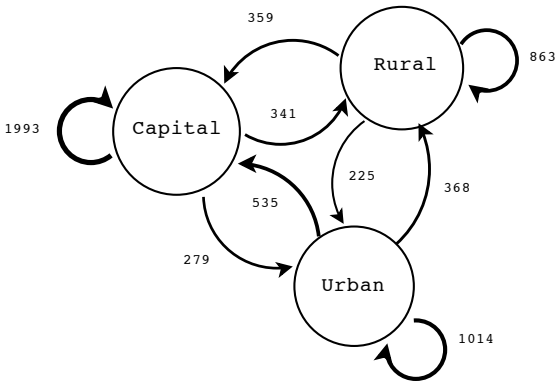

Fig. 1. Average Calling Volumes per Subscriber for the Capital, Urban, and Rural regions. (seconds)

B. Methodology

1) Segmenting Regions: We divide the 1.4 million subscribers into three categories based on geography: those living in the country's capital (600k), the other 11 urban towns (500k), and rural areas (300k). The regional ties based on calling volumes are shown in Figure 1.

2) Identifying Individuals: To establish an individual's regional label and weight, we identify the region where the individual spent the majority of time based on the cellular tower data for each week. For example, an individual who spends 3 weeks in the capital and then spends a week split between the capital and the rural area would have the regional label of 'Capital' with a weight of 0.875. The individual is then associated with the most probable region, and subsequently these weights are no longer used in this analysis.

IV. Individual Attributes

A. Socioeconomic Status via Calling Card Denominations

The vast majority of the subscribers are on pre-paid plans which necessitates the periodic purchase of airtime scratch cards, a ubiquitous commodity readily available in both the urban and rural areas of the county. Scratch cards are sold in a variety of denominations ranging from the equivalent of $0.25 to $20 dollars. We postulate that individuals who purchase higher denomination cards are more economically advantaged than individuals who purchase the same total amount of airtime incrementally using many, smaller denomination scratch cards. Figure 2 shows a one-way anova with a box corresponding to the scratch cards' lower quartile, median and upper quartile values. The lines extending beyond the box correspond to the 95% values. There is a notch in each box representing the uncertainty about the medians for box-to-box comparisons. Given our sample size (every mobile phone subscriber in the country) the notches are quite small and non-overlapping, indicative that the medians of the three

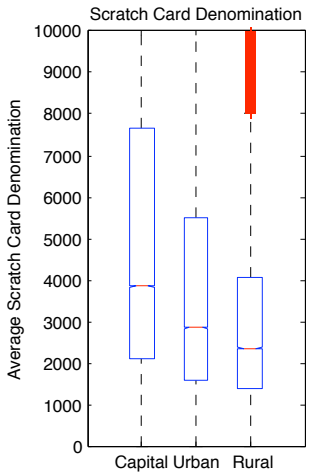

Fig. 2. Average Airtime Denominations Purchased in each Region (country currency)

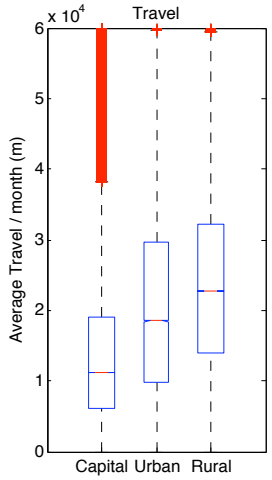

Fig. 3. Average Amount of Monthly Travel by Subscribers from each Region

groups differ at a 5% significance level. From the Figure 2, we can see that individuals living in the capital city use a card denomination that is almost twice as much as the the card denominations used within rural regions of the county. While this maps well to government census data about socioeconomic status levels within the capital, urban, and rural regions, it should be noted that we are unable to validate this apparent correlation.

B. Travel: Distances between Cellular Towers

We created a weekly metric for the amount of travel completed by an individual by calculating the maximum distance between the each week's set of used towers given by the equation below,

$$\sum_m \max_t \frac{\mathrm{d}(t_1,t_2)}{\mathrm{M}_p} \quad \forall t \in \mathrm{Set}_p(m)$$

where $\mathrm{d}(t_1,t_2)$ is the distance in meters between two towers (computed from the towers lat, lon), $\mathrm{Set}_p(m)$ is the set of towers used by user p during month m, and M_p represents the months when the user has been active.

It is clear from Figure 3 that individuals in the rural areas travel significantly more per month than individuals living in the cities. One reason for this simply could be due to the small potential distances that can be traveled within the capital and the much larger distances within rural areas.

C. Product Adoption: Airtime Sharing

Like many countries with a dominant pre-paid market, the operator launched a USSD airtime sharing application. To send airtime to another number, a users must type in a combination of the recipient's phone number, the hash key, and then amount of airtime to send to a special USSD short-code. Because this service is free, it was not significantly advertised and therefore the structure of this protocol presumably traveled by word of mouth. Despite the lack of publicity for this service, the majority of subscribers have used this service (58% in the capital, 57% in the urban regions, and 62% in the rural regions). While the rural regions have the highest percentage of users, Table I shows this region transfers the least amount of airtime.

TABLE I
AIRTIME TRANSFERS BETWEEN REGIONS (COUNTRY CURRENCY)

from\to	**Capital**		**Urban**		**Rural**	
	avg	*std*	*avg*	*std*	*avg*	*std*
Capital	515	(2630)	191	(2625)	83	(780)
Urban	121	(1268)	637	(4605)	118	(1385)
Rural	66	(824)	172	(2505)	268	(1591)

V. SOCIAL NETWORK ATTRIBUTES

There has been much theoretical as well as empirical work done in effort to quantify the role of cities in shaping personal networks [3], [4], [5], [9], [17], [16]. In this section we hope to add to this literature by quantifying the communication rates and topological properties associated with personal networks from mobile phone subscribers in both urban and rural regions.

A. Frequency and Volume

As theorized previously, we are able to validate that individuals living in urban areas tend to communicate almost 50% more than individuals living in rural areas. Figure 4 shows the distributions associated with the average outgoing call volume and frequencies for each of the three regions while Figure 1 shows that urban regions tend to make more inter-regional calls than they receive.

B. Degree and Average Volume per Degree

Previous work has theorized that people in rural areas have relatively strong ties to fewer people, whereas individuals in urban areas tend to have more, but weaker ties [3], [4], [5], [6], [9]. We find that this is also the case in our data. Figure 5 and Figure 6 show that individuals living in rural areas have distinctly lower degree than those living in the capital (median degree of 109 vs. 175). However we also find that while these individuals in rural areas may have fewer ties, they have higher

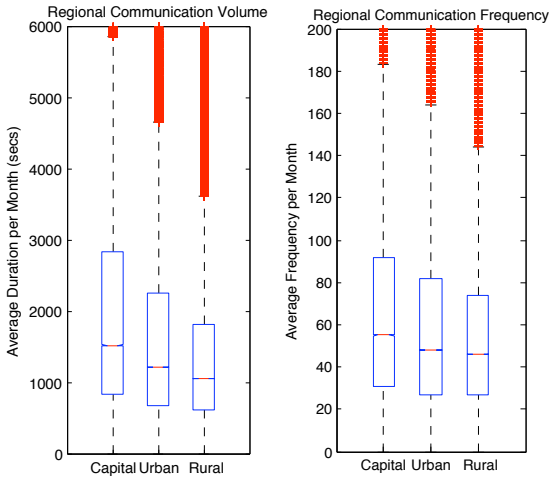

Fig. 4. Average Outgoing Call Volume and Frequency

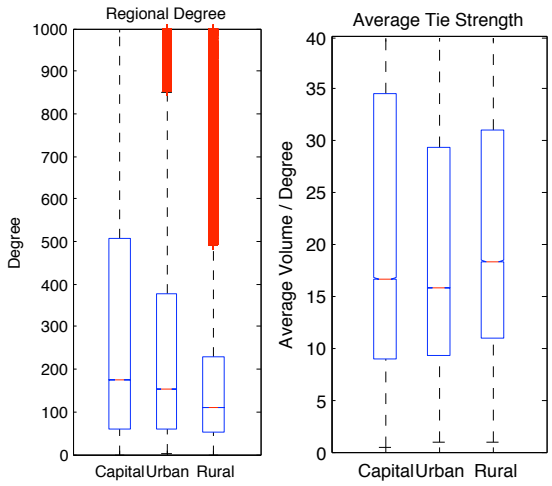

Fig. 5. Average Degree and Tie Strength

average tie strength (defined as the total volume / degree). This validates and quantified previous qualitative theory about the role of the city in personal networks.

C. Alter Attributes

In this section we not only look at degree and volume of an individual, but also take into account attributes of the individual's personal network. We will study the intra and inter regional communication patterns, calculate clustering and egodensity metrics on these personal networks, and measure the number of an individual's communication partners who also communicate with each other.

1) Local vs. Long Distance Edges: Table II shows the call behaviors of each of the three regions. While the majority of phone calls stay within the same region, the variance in each category is quite high, but averages are close to being symmetric.

2) Egodensity and Clustering: We select N=5000 random individuals to compute more computational intensive network metrics including clustering coeffient and egodensity. We

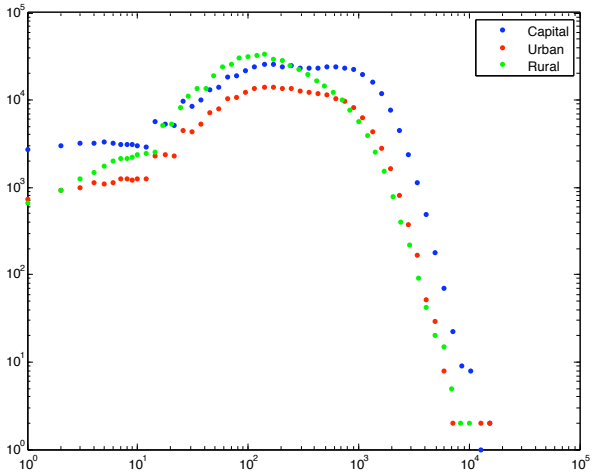

Fig. 6. Degree Distributions

TABLE II
AVERAGE CALL VOLUMES BETWEEN REGIONS PER SUBSCRIBER (SECONDS)

from\to	**Capital**		**Urban**		**Rural**	
	avg	*std*	*avg*	*std*	*avg*	*std*
Capital	1993	(4619)	279	(834)	341	(787)
Urban	535	(1524)	1014	(2082)	368	(763)
Rural	359	(930)	225	(566)	863	(1267)

define egodensity as the percentage of existing edges within the egocentric network over the total possible number of edges as,

$$\text{egodensity} = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_v \text{edges}(n,v)}{k_n(k_n-1)} \quad (1)$$

where edges(n,v) is a binary function that equals one if there is an edge between node n and v and zero otherwise. The sum is averaged over all nodes in the sample and k_n is the degree of node n.

Results are shown in Figure 7, demonstrating that rural and urban networks are denser and more clustered than those in the large city, as we may also have anticipated from the general considerations above.

VI. Behavioral Persistence

Most data used for social research (including those generated by mobile phones) tend to analyze static, behavioral snapshots. However, longitudinal data are essential to discriminate between cause and effect in behavior data. For example, in some ongoing research on the effect cities have on their inhabitants' social networks, we find that individuals who live in cities tend to have on average different types of social networks than those who live in rural areas [9]. However, a legitimate critique of this result is that the original question of causality has gone unanswered.

Specifically two simple mutually exclusive hypotheses could account for these effects: 1) differential selection and 2)

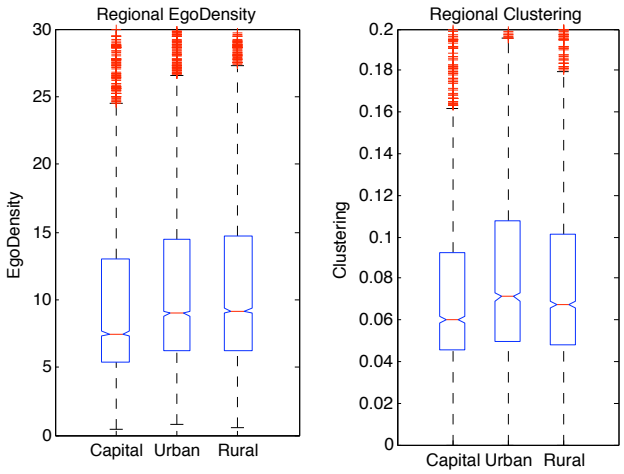

Fig. 7. Personal Network Egodensity and Clustering Coefficient (5000 samples per region)

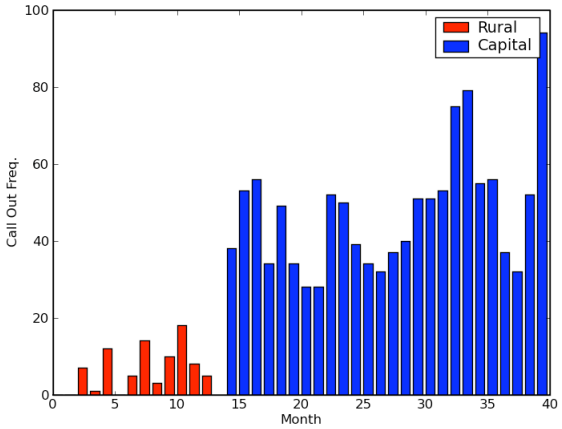

Fig. 8. The call frequency of a typical individual living in a rural area and then moving to the capital.

behavioral adaptation. Under the differential selection hypothesis, rural and urban environments have different distributions of behavior. This hypothesis asserts that individuals do not change their behavior as they migrate, but rural individuals with larger and more diverse social networks may tend to move to the large city, and vice versa for urban dwellers with small networks. Under behavioral adaptation, individuals align their behavior with what is typical for their social environment, responding to the greater number of more diverse social opportunities of a large city by expanding their social networks, and reducing them accordingly when moving to rural areas.

With the current 'snap-shot' data, we can not tell whether the city attracts individuals who already have a signature social network, or whether indeed the city itself influences the network of its inhabitants. To obtain a better answer to this question it is necessary to have longitudinal data. Now that we have over four years of data on every mobile phone subscriber in the country, we can identify individuals who live in rural areas during year one, and then move to urban areas in year two. By comparing their before and after social networks we can get a better idea of the effect of the city. Indeed with several years of data, we can also learn if these individuals maintain new relationships created by the urban area if they move back to their rural home.

We investigated the persistence of average call frequencies under circumstances when individual migrate, either from rural areas to the capital or in the opposite direction. Results are summarized in Figure 8, 9, and 10. We find that while migration increases total call volume, in most cases call volume to the former region decreases. However, call frequency to the new region increases, especially when the movement is from a rural region to the capital as shown in Figure 9.

We find results in support of the behavioral adaptation hypothesis. Individuals migrating to the city subsequently increase their call frequencies matching the behavior expected

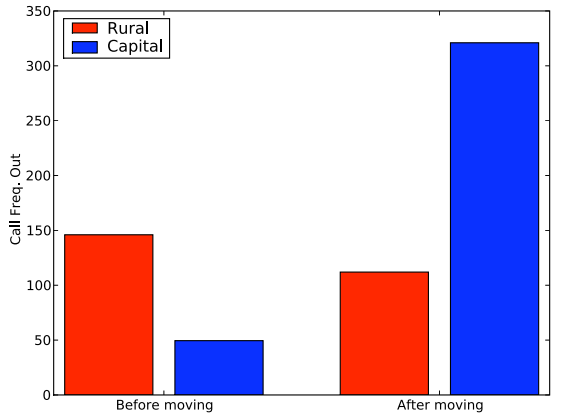

Fig. 9. The average call frequencies to rural regions and the capital of 50 sampled individuals who were living in a rural region and subsequently moved to the capital.

for a typical urban dweller. Importantly the result is also true in the opposite direction: individuals moving to rural areas from the city undergo a reduction in communication commensurate with the average for the rural areas to which they have gone.

VII. Discussion

In this paper we have proposed the use of cell phone usage data to test, elaborate and quantify classical hypotheses in sociology, social psychology and economics about behavioral changes and human and social adaptations as a result of life in large cities versus smaller urban areas and rural settings. We have argued that this type of data can now supply statistical coverage of the majority of the population, albeit through technologies that allow us to measure specific quantities that are correlates of cost, rhythms of life, and the dynamics and structure of social networks.

We have found support, and quantified here for the first time on a large scale, for arguments for the diversification

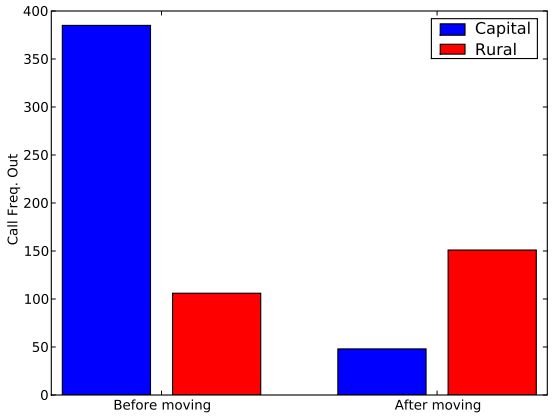

Fig. 10. The average call frequencies to the capital and rural regions of 50 sampled individuals who were living in the capital and subsequently moved to a rural region.

and growth of personal networks as individuals live or move to large urban areas. We have also found evidence that this growth results in a optimization process whereby the burden of maintaining large number of social contacts is partially mitigated by the fact that most of these contact are weak, taking up less of the individual's time. At the same time we characterize several other personal attributes of mobile phone users and their geographic disparity.

Finally, we were able to test statistically two alternative hypothesis for the origin of these effects, namely whether individuals change behavior to conform with their social environment (behavioral adaptation) or instead migrate to realize their preferences for larger and more intense social environments of the large city or a smaller number of stronger links characteristic of rural areas (differential selection). We found strong support for behavior adaptation over differential selection, although future work in under way to further test the viability of these two alternative scenarios.

A. Causality

There remain several confounding factors that limit the scope by which we can compare the effects of rural and urban societies. Perhaps the most significant are the socioeconomic discrepancies between the different regions, which could certainly explain many of the results including the increased communication and travel. Establishing causal relationships from urban environments will remain elusive until more rigorous experiments are designed and performed.

B. Future Work

While we have completed a preliminary comparison of urban and rural communities, there is still much more to be done. We intend to study in greater temporal detail the migration behavioral changes of individuals, and the ties they maintain through these periods of change with their point of origin and the formation of new ones in their destination. We also intend to study the diversity of several personal attributes in urban and rural areas and their genesis and evolution, e.g. as a new product is introduced and adopted, and whether these processes are facilitated by underlying social networks of communication.

VIII. Conclusion

This paper represents an initial analysis of how mobile phone data can provide comparative insight into human and social behavior in urban and rural communities. We have tested, confirmed and quantified classical hypothesis in sociology, social psychology and economics that urbanization leads to increased communication, and present a methodology for inferring socioeconomic status based on airtime top-up denominations. We have also confirmed hypothesis for behavioral adaptation of individuals based on changes in their patterns of communication to increase the similarity with their new social environment. We believe more detailed analysis of this and other data sets will shed additional light not only on the structural changes in social and human behavior between rural areas and large cities but also on the principles and mechanisms that enable these changes. Such results will advance the conceptual framework in the social sciences and economics and may result in new approaches to public policy.

Acknowledgment

The authors would like to acknowledge Vincent Blondel, our anonymous reviewers, as well as the Santa Fe Institute.

References

[1] J.J. Shaughnessy, E. B. Zechmeister, and J. S. Zechmeister, *Research Methods in Psychology*, Higher Education, New York, NY, 2006.
[2] C. Moser, and G. Kalton, *Survey Methods in Social Investigation*, Dartmouth Publishing Co Ltd, Darthmouth, NH, 1985.
[3] F. Tönnies, (1963) *Community and Society* Harper & Row, New York, NY, Gemeinschaft und Gemeinschaft, originally published in German, 1887.
[4] G. Simmel, (1964) The Metropolis and Mental Life, p 409-24 in *The Sociology of George Simmel*, ed. Wolff, K., Free Press, New York, NY; originally published in German, 1903.
[5] Wirth, L. Urbanism as a way of life *Am. J. Sociol.* 44:1-24, 1938.
[6] S. Milgram, The experience of Living in Cities, *Science*, 167: 1461-1468 1970.
[7] Macionis, J. J. and Parillo, V. N. (1998) *Cities and Urban Life*. (Pearson Education Inc., New York, NY).
[8] E. Glaeser Is There a New Urbanism? The Growth of U.S. Cities in the 1990s, *Cityscape* 1: 9-47 (1994).
[9] C. Fischer, *To Dwell among Friends: Personal Networks in Town and City*, Chicago, United States: University of Chicago Press, 1982.
[10] F. Hollinger and M. Haller, "Kinship and social networks in modern societies: a cross-cultural comparison among seven nations", *European Sociological Review* 6:2, pp. 103-124, 1990.
[11] M. Gonzalez, C. Hidalgo, and L. A. Barabasi, "Understanding individual human mobility patterns", *Nature* 453, pp. 779-782, 2008.
[12] J.Onnela, J.Saramaki, J.Hyvonen, G.Szabo, D.Lazer, K.Kaski, J.Kertesz, and A.-L.Barabasi, "Structure and tie strengths in mobile communication networks", *Proceedings of the National Academy of Sciences* 104, (2007), pp.7332–7336.
[13] N. Eagle, "Behavioral Inference Across Cultures: Using Telephones as a Cultural Lens", *IEEE Intelligent Systems*, 23:4, 62-64, 2008.
[14] G. Szabo and L. A. Barabasi, "Network effects in service usage", *arXiv:physics/0611177*, 2006.
[15] S. Hill, F. Provost, and C. Volinsky, "Network-Based Marketing: Identifying Likely Adopters via Consumer Networks", *Statistical Science*, 21:2, 256-276, 2006.

[16] F. Calabrese and C. Ratti "Real Time Rome", *Networks and Communication Studies - Official Journal of the IGU's Geography of Information Society Commission*, 20:3, 247-258, 2006

[17] L. M. A. Bettencourt, J. Lobo, and D. Strumsky Invention in the city: increasing return to patenting as a scaling function of metropolitan size, *Research Policy*, 36: 107120 (2007) .