LBSN Data and the Social Butterfly Effect (Vision Paper)

Clio Andris
The Pennsylvania State University
302 Walker Building
University Park, PA 16802
+1 814 865-1175
clio@psu.edu

ABSTRACT

LBSN data are well-suited for research questions and perspectives on social or spatial phenomena. Researchers often subset large LBSN datasets into different social networks (using snowball sampling), temporal or spatial granularities, to test for statistical patterns. Yet, researchers lack a way to examine how human interpersonal behavior results in digital traces of geolocated social events, although macro global flows of movement and communication are built from micro individual human intentions.

To help navigate between the individual mind and the resultant big LBSN data that researchers use to understand society and space, I list a 14-tier scale of connectivity typologies. Each step can provide different a perspective of a single LBSN dataset. This scale can illustrate how perturbations at one level affect another level. E.g. How will reported escalating rates of autism affect the future network of connectivity between global cities? Will a change in migration policy strain emotional ties between an international family? The scale allows us to track changes at different levels between micro-, meso- and macro-scale social-spatial phenomena in a computationally-friendly way.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics, Psychology, Sociology

General Terms

Economics, Theory.

Keywords

Interpersonal Relationships, GIS, Scaling, Social Networks, Travel, Telecommunications, Human Behavior, Complexity.

1. INTRODUCTION

Lorenz posed the metaphorical question, *Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?* to illustrate how a small, seemingly-unimportant change on a micro scale can cause changes on the macro scale [1]. This metaphor extends to other domains. When a butterfly flaps its wings in a relationship (i.e. family, professional, friend, romantic, etc.) the hurricane that occurs is not confined to the relationship, nor to the larger social network, but to the built environment as well.

LBSN'15, November 03-06 2015, Bellevue, WA, USA

ACM 978-1-4503-3975-9/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2830657.2830658

Data that evidence geolocated relationships are called location-based social networks (LBSN), and are increasingly used by computational scientists for sociological and urban applications. LBSN data are multi-faceted. The same data set can be used to explore spatial interaction between an agent and his wife [2], or continent-to-continent calling patterns [3]. This versatility shows that social/spatial behavior is connected in some way; it all belongs to the same story. Researchers should leverage the ability to unearth individual-scale decision-making [2] and consider how this behavior extrapolates (or does not) to the big data that LBSN researchers mine. They should incorporate the micro: human cognition (specifically the natural drive to form relationships and fill personal, social needs) and the macro: large-scale movement and telecommunications patterns across geographic space, and the meso-scale intermediary snapshots of social ties and geography.

1.1 Macro and Micro in LBSN

How do researchers use LBSN data to view how changes in the social mind will affect changes in large scale human connectivity? Successful findings in micro-to-macro connectivity in LBSN type data include explaining how a living cell's properties are mimicked in cities [4] and how urbanites' intellectual properties scale with city size [5]. Further research shows that individuals do not want to live in spatially segregated cities, but choose residences that reinforce segregation [6]. In agent-based traffic models, individual agents are programmed with simple micro-rules to facilitate their individual quick passage, but eventually produce macro-traffic [7]. In chain migration, an agent will relocate to a new place in order to earn more money and search for better opportunities, and once capital is built, other families will join [8]. As a result, entire settlements have been transferred to new places depending on the trust and leadership of very few people. A single LBSN dataset can scale naturally from one record's intimate story to big data. In a classic example of a single laboratory's treatment of a large LBSN dataset of AT&T's New York City-based global telephony, researchers focused on different narratives. One visited New York City to conduct interviews of immigrant telecommunications [3] while another used eigenvector decomposition to create an unsupervised classification of New York neighborhoods [9]. A third member visualized the data for New York's Museum of Modern Art [mentioned in 3] and a fourth examined day-calling and night-calling patterns.

1.2 The Behavioral Variable

LBSN researchers are experts at subsetting and aggregating data by variables, most commonly by spatial scope and granularity, temporal scope and granularity. They also make important decisions about summary statistics: such as call duration, number of trips, etc. These data are often treated the same as if they were of weather or thermodynamics. The data have interesting

numerical variables that can be rearranged and correlated. For instance, data on sequential GPS traces are mined for their tortuosity, length and turn patterns—sometimes more creatively with the intersection and proximity of other geographic features. In order to push the computational and mathematical field, it sometimes seems prudent to focus attention away from what the data represents, as to focus on algorithms and performance.

Such communities can broaden or re-focus their techniques and technology in a way that evidences how LBSN data is actually a representation of the real world and real people. There are indeed rich stories behind each dataset row. To the computer scientist, the term *story* can equate to more data. The first step, however, is to provide LBSN researchers with methods of aggregating and disaggregating data by human social behavior granularity.

The purpose of this article is to present a scale tailored to the big-data era that can still offer insight into the invisible voices behind LBSN flat files: human social intentions. A set of intermediary steps can explain how micro cognition and macro socio-geographic processes are intertwined. The scale guides the LBSN researcher from the individual mind (A1) to the masses of spatial connectivity data mined daily by researchers worldwide (G1 and G2). The goals are to analyze the symbiotic connection of interpersonal relationships and geographic space at different scales and levels of detail; describe how social relationships (friendships, co-workers, family, etc.) guide spatial connectivity, the shaping of place, settlement and globalization; chart ripple effects across the scale, such as how a change in one's ethnic stereotypes and affects his or her activity paths. The scale can help describe the role of place and environmental features (infrastructure, natural features, and social divisions) in enabling or stress on personal relationships and vice versa.

2. THE SOCIAL BUTTERFLY SCALE

This scale illustrates the steps between the individual mind and resultant global flows via 14 categories of phenomena, divided into groups A-G with general examples following in italics (also represented in Figure 1). Note that there is a vast imbalance of the magnitude of research dedicated to each category. Research in groups D and F have been approached less than the remainder due in part to the newness of data and methods, and privacy issues. I list a few examples of theoretical work in each category. Instead of citing modern LBSN applications at various tiers, I chose classic illustrations when possible to expose LBSN researchers to interdisciplinary resources and early formalizations that underlie these configurations. (In the following section, level and group are used interchangeably).

A SCALE FOR NAVIGATING FROM INDIVIDUAL BEHAVIOR TO LARGE-SCALE SPATIAL PATTERNS[1]

- **A1** Neurological capacity and cognition: *Agent (i.e. an 'ego') can communicate with an 'alter' (i.e. an agent who is friends with the ego)* [10].
- **A2/B1** Social cognition and behavior: *Agent sees an alter as a potential friend* [11].
- **B2/C1** Dyadic connections: *Agent is friends with alter* [12].
- **C2** Assemblage of dyadic relationships, groups (social capital): *Agent has siblings, parents, children, classmates, co-workers, and friends* [13].
- **C3** Configuration of dyadic relationships (social network): *Agent's siblings and parents form family clique (i.e. fully connected social network). Agent has friends who are acquainted with each other* [14].
- **D1** Spatial distribution of dyadic connections: *Agent's friend lives in city x* [15].
- **D2** Spatial distribution of social capital: *Agent has friends in cities x, y and z* [16].
- **D3** Spatial distribution of social networks and institutions *Agent has a clique of 6 friends in city z and 2 friends in city y. Agent belongs to a club in city z* [17].
- **E1** Individual spatial communication patterns: *Agent calls friend in city x* [18].
- **E2** Individual movement patterns & activity spaces (mobility & accessibility): *Agent takes train to city y* [19].
- **F1** Socially-linked spatial communication patterns: *Agent sends group e-mail to friends in various locales. Agent's mother calls a friend in city x* [20].
- **F2** Socially-linked movement patterns: *Agent travels via airplane to visit clique of college friends in city z. Agent's sister vacations in city x* [21].
- **G1** Large-scale spatial communication patterns: *City x calls city y 10,000 times per month* [22].
- **G2** Large-scale movement patterns & human land use patterns (large-scale mobility & accessibility): *7,000 agents migrate from city x to city y each year* [23].

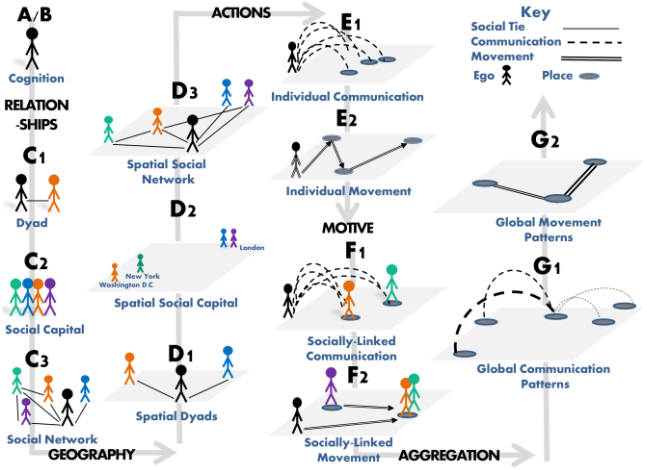

Figure 1. A pictorial representation of a scale for navigating from individual behavior to large-scale spatial patterns.

The following is a synthetic social situation used to illustrate how decisions are determined in the mind (almost instantaneously) with both social and spatial variables at hand. A woman (*W*) says: "I am angry with my spouse. Instead, I will spend time with my sister." When she looks to another alter for social fulfillment, her spatial behavior will shift: "I will not travel with my spouse. I will visit my sister in Germany." Her perception of whether the spouse

[1] Example SQL queries to retrieve data configured at each level of the scale are available at http://personal.psu.edu/cma24/SQL

(*M*) or sister (*S*) will satisfy her interpersonal needs will determine her trajectory. If her spouse writes her a card, she may make amends, and will cancel her ticket to Germany, or they will travel together. This results in a change in a large origin-destination matrix of flights. For a more general example see Milgram's 1967 letter experiment [20] and modern derivations [24], where all flows can be classified under case F1 but can be aggregated and disaggregated to different levels of the scale.

2.1 A & B: Individual Cognition

The scale begins by acknowledging a factor that will drive large-scale connectivity: the agent's capacity for socialization (A1). In group B, the agent creates relationships and displays social behavior. Groups A and B are not the focal point of this research, but their properties reflect the initial configuration that is integral for showing how social behavior emerges. *W*'s perception of others and her psychology is a prior consideration to her behavior. *W* has amassed contacts as an effect of her (and her alters') social cognition behavior. The argument between *W* and *M* is tempered by these factors [2]. In the future, perhaps the researcher can encode an agent's gut feelings, mental models, reactions, perspectives of danger, identity, tipping points etc. in group B to examine the byproducts of his or her behavior in geography. Do certain personality types frequent certain locales?

2.2 C: Social Relationships

Group C shows digital proof of the agent's set of contacts via phone calls or online media, etc. In the synthetic example, *W* conjures of a list of her alters (C1). Of these, *W* may consider which alters would help calm her emotional state, that is, provide social capital (C2). Following, *W* considers which *set* of alters or groups of alters may offer the most support by envisioning the configuration of her social capital as a social network (C3). Group C shows that *W* has 40 alters and that *W*'s best friend has not called recently, changing their relationship within a larger net of social ties. It reveals that *W* calls her sister *S* and mother every other day. *W* chooses a family member who is embedded in similar social circles or cliques to help her. *W*'s chooses *S* because *S* is less likely to spread information about her situation to other family members (known via the social network). *W* avoids members with strong ties to *M*. The rich field of social network analysis (SNA) is essential for showing these dynamics.

For computational implementation, the dyad is the recommended atomic unit (i.e. primary key, unique identifier) that links the built environment and human social behavior because it expresses social and spatial intentionality (i.e. reasons for acting) given fundamental social needs, and serves as an instant origin-destination flow, since two people separate across geographies. This is recommended over the individual as the unique ID, whose behavior does not evidence clear social intentionality or provide a view of his or her ability to socialize. Understandably, the dyad requires more computation power and a results in a larger dataset.

2.3 D: Spatial Distribution of Relationships

In group D, geographic space becomes a variable. *W's* alters are geolocated, and thus counted, assembled and clustered within geographic space to show a social topology. She may have a wealth of contacts in a single location or few contacts in distributed locations, etc. *W* may consider her relationships with others in combination with their location (D1) and her social capital's utility depending on their locations (D2): revealing the economics of choosing a helpful, faraway contact, or a less helpful nearby contact. She considers the configuration of this social network across the region (D3) as well, as nodes and edges within geographic space and can be used to choose a location of a support member with, for instance, high SN centrality. From here, *W* can determine the costs and benefits of contacting or meeting one of her alters. Whereas SNA is essential for group C, GIS is essential for group D.

2.4 E: Individual Actions

In group E, geographic activity is revealed through evidence of behavior: the agent telecommunicates (E1) (calls, SMSs, sends e-mails, letters, online messages) and moves (E2) (walks, takes a train, migrates). These actions are visible through the digital records LBSN researchers often mine for patterns. For example, *W* tries to avoid *M*'s typical movement patterns and chooses to visit a place to where she has not recently traveled (E2). *W* calls Germany to coordinate her plans. It is important to distinguish between telecommunicating (E1) and movement (E2) because each offers different payoffs and require the built environment to play different supporting roles—yet both are often correlated [12].

2.5 F: Socially-Linked Actions

Group F may be the most complex configuration of variables modeled using LBSN data. Group F combines level D with level E data to show whether the origin-destination trips exhibited in layer E spatially correlate with the locations where agents have contacts (as described in level D). It can also be understood as a dynamic representation of the 'set-stage' of D, now full with action and behavior.

Computationally, communications or travel traces can be overlaid in GIS with the social topology shown in level D to determine which traces (E) spatially overlap with a contact (D), which may provide evidence of a socially-driven action. Of course, the LBSN data cannot prove that *W* travels to meet an alter individual, but we understand that *W* (sub)consciously is likely to prefer environments near trusted contacts. *W*'s call to Germany is now linked to her sister *S*, who makes most of her calls from Germany. A non-socially driven action, *W*'s a call to an automated weather service or her solo trip to a survey location represent (E) traces that do not overlap with any (D) data—or do by happenstance.

2.6 G: Large-Scale Communication and Movement Patterns

In group G, the agent becomes one of many agents linked to the masses. These traces include communication patterns (G1) and movement patterns (G2), as aggregated to geographic origins and destinations. The values associated with these origins and destinations (i.e. edge weights) are often computed based on number of telecommunications events or duration, or number of trips. Like group E, these data are typical for LBSN researchers. *W*'s decision to visit *S* in Germany, remain at home with *M*, or go with *M* to visit *S* in Germany will result in 1, 0 or 2, respectively, additions to a researcher's aggregate O-D international flight matrix. Holding financial and national diplomatic scenarios constant, these values are determined by emotions, dialog, interpersonal ties and the locations of these ties.

3. DISCUSSION AND CONCLUSION

In summary, a 14-tier scale of connectivity typologies is outlined to help guide data analysts between understanding the individual and the big LBSN data he/she creates. Different derivation can be

employed on a single LBSN dataset to provide multiple perspectives on the human behavior that creates LBSN data.

The reasons for some large-scale LBSN behavior may seem clear, such as migration after the infestation of the boll weevil and potato blight, or U.S. retirees moving to Florida. Or, a downtown concert may draw city residents. But in these examples, interpersonal relationships are still necessary: To where did emigrants travel after the infestation? Was a retirement community more attractive because of existing friendships? Did the city concert draw individuals, small families, or romantic couples? Incorporating ideas from psychology, sociology, geography and spatial economics may lead to more fruitful hypotheses and rewarding research. Importantly, seemingly-individualized, unique, unclassifiable flows (i.e. not refugee evacuations) are catalyzed by similar personal reasons: financial (travel to work), social support (attending a wedding or recital), and others: romance, hobby interests, and errands. This may be where the micro-to-macro viewpoint reveals its power. As in complex systems, a major challenge to using this tool is incurring the burden of proof that one system component affects another component (and is not just correlated with spatial or temporal changes).

From this discussion, I encourage the computer scientist and data engineer to derive methods that can account for human behavior. In terms of operationalization, the social butterfly scale can be implemented as a SQL (or graph based) query system, so that certain layers can be retrieved from the data[1]. Researchers already use operationalized systems where the social network and spatial variables are selected in tandem, but the aggregations provided in this scale allow the researcher to assemble and disassemble data methodically. Its linear directionality gives the researcher (and student) a set of empirical steps.

ACKNOWLEDGMENT

Thanks to Sara Cavallo for assistance with the communication of these ideas.

4. REFERENCES

[1] Lorenz, E. 1972. *Predictability: Does the Flap of a Butterfly's Wings in Brazil Set Off a Tornado in Texas?* Address at the 139th Annual Meeting of the Am. Assoc. Adv. Sci. Boston, MA, December 29, 1972.

[2] Gottman, J., Murrey, J., Swanson, C., Tyson, R., Swanson, K., 2005. *The Mathematics of Marriage: Dynamic Nonlinear Models.* MIT Press, Cambridge, MA.

[3] Rojas, F. 2010. *New York Talk Exchange: Transnational Telecommunications and Migration in a Global City.* Doctoral Thesis. UMI Order Number: UMI Order No. 0823186. Massachusetts Institute of Technology.

[4] Couclelis, H. 1985. Cellular worlds: a framework for modeling micro-macro dynamics. *Env. Plan. A.* 17, 5, 585-596.

[5] Bettencourt, L., Lobo, J. and Strumsky, D. 2007. Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. *Research Pol.* 36, 1, 107-120.

[6] Schelling, T. 1978. *Micromotives and Macrobehavior.* Norton, New York.

[7] Cetin, N., Burri, A. and Nagel, K. 2003. A large-scale agent-based traffic microsimulation based on queue model. In *Proceedings of the 3rd Swiss Transport Research Conference* (Monte Verita, Ascona, Switzerland, March 19 - 21, 2003). STRC '03.

[8] Pedraza, S. 1991. Women and migration: The social consequences of gender. *Annu. Rev. Sociol.* 17, 303-325.

[9] Reades, J., Calabrese, F. and Ratti, C. 2009. Eigenplaces: analysing cities using the space- time structure of the mobile phone network. *Env. Plan. B.* 36, 5, 824-836.

[10] Minksy, M. 2006. *The Emotion Machine*. MIT Press, Cambridge, MA.

[11] Kunda, Z. 1999. *Social Cognition*. MIT Press, Cambridge, MA.

[12] Stafford, L. 2004. *Maintaining Long-Distance and Cross-Residential Relationships*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

[13] Fernandez, R., Castilla, E. and Moore, P. 2000. Social capital at work: networks and employment at a phone center. *Am. J. Sociol.* 105, 5, 1288-1356.

[14] Bearman, P. S., Moody, J. and Stovel, K. 2004. Chains of affection: The structure of adolescent romantic and sexual networks. *Am. J. Sociol.* 110, 1, 44-91.

[15] Fischer, C. 1982. *To Dwell Among Friends: Personal Networks in Town and City.* University of Chicago Press, Chicago, IL.

[16] Onnela, J. P., Arbesman, S., González, M. C., Barabási, A. L., and Christakis, N. A. 2011. Geographic constraints on social network groups. *PLoS ONE*, 6, 4, e16939.

[17] Hampton, K. and Wellman, B. 2003. Neighboring in Netville: How the Internet supports community and social capital in a wired suburb. *City Comm.* 2, 4, 277-311.

[18] Kwan, M. 1999. Gender and individual access to urban opportunities: a study using space-time measures. *Prof. Geogr.* 51, 211-227.

[19] Carrasco, J. and Miller, E. 2006. Exploring the propensity to perform social activities: a social network approach. *Trans.* 33, 5, 463-480.

[20] Milgram, S. 1967. The small world problem. *Psychol. Today. 2*, 1, 60-67.

[21] Radil, S., Flint, C. and Tita, G. 2010. Spatializing social networks: using social network analysis to investigate geographies of gang rivalry, territoriality, and violence in Los Angeles. *Ann. Assoc. Am. Geogr.* 100, 307-326.

[22] Rietveld, P. and Janssen, L. 1990. Telephone calls and communication barriers. *Ann. Reg. Sci.* 24, 4, 307-318.

[23] Limtanakool, N., Schwanen, T. and Dijst, M. 2009. Developments in the Dutch urban system on the basis of flows. *Reg. Stud.* 43, 2, 179-196.

[24] Liben-Nowell, D. and Kleinberg, J. 2008. Tracing information flow on a global scale using internet chain-letter data. *Proc. Nat. Acad. Sci. U.S.A.* 105, 12, 4633-4638.