Individual Behavior and Social Influence in Online Social Systems

Manos Papagelis
University of Toronto
papaggel@cs.toronto.edu

Vanessa Murdock
Yahoo! Research
vmurdock@yahoo-inc.com

Roelof van Zwol
Yahoo! Research
roelof@yahoo-inc.com

ABSTRACT

The capacity to collect and analyze the actions of individuals in online social systems at minute-by-minute time granularity offers new perspectives on collective human behavior research. Macroscopic analysis of massive datasets raises interesting observations of patterns in online social processes. But working at a large scale has its own limitations, since it typically doesn't allow for interpretations on a microscopic level. We examine how different types of individual behavior affect the decisions of friends in a network. We begin with the problem of detecting social influence in a social system. Then we investigate the causality between individual behavior and social influence by observing the diffusion of an innovation among social peers. Are more active users more influential? Are more credible users more influential? Bridging this gap and finding points where the macroscopic and microscopic worlds converge contributes to better interpretations of the mechanisms of spreading of ideas and behaviors in networks and offer design opportunities for online social systems.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information Networks*; J.4 [**Computer Applications**]: Social and Behavioral Science

General Terms

Algorithms, Human Factors, Experimentation

Keywords

Social Networks, Behavior, Influence, Diffusion, Geotagging

1. INTRODUCTION

Over the last several years the Web witnessed a prolific growth largely due to the changing trends in the use of Web 2.0 technology that aims to enhance interconnectivity, self-expression and information sharing. These trends have led to the development and evolution of virtual communities and services, such as social networking sites, photo and video sharing services, blogs, wikis, and folksonomies, but also to the creation of voluminous user-generated data in the form of text, images, videos and more. Even though these new social media change the way we communicate, the underlying social processes remain governed by long-standing principles of human behavior and social interaction.

To date, analysis of human behavior and interaction has been limited to a small number of self-reporting individuals, due to practical issues. As such, models and algorithms typically have been based on a small number of temporary snapshots of the network structure and data. On the other hand, social processes that take place in online spaces can be monitored at unprecedented levels of scale and resolution, producing massive datasets. The capacity to collect and analyze the actions of millions of individuals at a minute-by-minute time granularity offers new perspectives on collective human behavior.

Macroscopic analyses of massive datasets raise interesting observations of temporal patterns of communication within social systems. The way we form connections, imitate actions, follow suggestions, influence others in forming opinions or making decisions, as well as the spread of behavior and the information flow in groups, describe dynamics and phenomena of everyday life, now expressed in an online setting. However, working at a large scale has its own limitations. Monitoring social activity in an aggregate fashion typically does not allow for interpretations on a microscopic level. Questions of how influential an individual is and what makes her more (or less) influential among her neighbors to a large extent remain disputed and open. Leveraging the interplay between macroscopic and microscopic worlds of online social processes and being able to find points of convergence between them is a challenge [19] and the main focus of this paper.

In this work, we examine how different types of individual behavior affect the decisions of friends in a network. We begin with the problem of detecting social influence in a social system. In the presence of social influence, an idea, behavior norm, or product can diffuse through the social network like an epidemic. Thus being able to identify situations where social influence prevails in a social system is important. Our analysis of social influence is based on the diffusion of a technological innovation. Then, we focus on two types of behavior expression, one that characterizes the *quantity* property (i.e., how often) and one that characterizes the *quality* property (i.e., how well) of the expressed behavior.

Detecting social influence derives from macroscopic analysis of aggregated data, while characterization of user behavior derives from microscopic analysis of user-specific actions. This setting allows to investigate causality between individual behavior and social influence in a principal way. In particular, we observe the diffusion of an innovation among social peers and try to identify what is the effect of particular behaviors in the social influence that individuals

HT'11, June 6–9, 2011, Eindhoven, The Netherlands.

exert to their network. This raises several research questions: Are more active users more influential? Are more accurate users more influential? How do we obtain and evaluate this information?

Our analysis of social influence is based on the adoption of the *geotagging innovation* in Flickr[1]. *Geotagging* refers to the process of adding geographical identification metadata to uploaded photographs, usually consisting of latitude and longitude coordinates, by placing them on a map. In our analysis, we need to look closer to individual actions and monitor how systematically and how accurately they use the innovation. We carefully design the experiments around Flickr and its geotagging innovation as it provides information on user actions in an adequate degree of detail for carrying out such research. Given this environment we make the following contributions:

- We present a method that given a social network and a log of user actions over a time period, detects and quantifies the occurrence of social influence among peers. We apply this method to show that adopting and using the geotagging innovation in Flickr does not happen randomly, but can to a large extent be attributed to social influence among users in the network (Section 3).
- We present a method that allows to characterize the *quality* property of a user's behavior by evaluating the accuracy with which they use the innovation - a measure of *user credibility*. (Section 4).
- We hypothesize there is a relation between a user's social influence and specific expressions of individual behavior and design experiments to test the hypotheses (Section 5). Our findings reveal an essential gap on the effect that these types of behavior have on influencing other people in their network and suggest design opportunities for online social systems.

The basic issue which we try to deal with in this research is the way in which an individual's choices depend on what other people do. Our work suggests an experimental framework of investigating causality of individual behavior and social influence in a network, while the methods we describe make a relatively small number of assumptions about the data, are general and can probably be easily adapted to other analysis of social systems.

2. RELATED WORK

The process of ideas and practices spreading through a population contagiously, with the dynamics of an epidemic, has long been of interest in the social sciences [38]. Its systematic study developed in the middle of the 20th century into an area of sociology known as the *diffusion of innovations* [32]. The theory of diffusion of innovations examines the effect of word-of-mouth information sharing and explores the role of social networks in influencing the spread of new ideas and behaviors. At a particular point in time, some nodes in the network become aware of new ideas, technologies, fads, rumors, or gossip, and they have the potential to pass them on to their friends and colleagues, causing the resulting behavior to cascade through the network. Models of diffusion of innovations in a social network have since been considered in many disciplines including mathematical sociology [12, 13], mathematical epidemiology, viral marketing, and game theory (see Kleinberg [11, 18] and the references therein). Diffusion of innovations is usually related to the problem of finding influential nodes in a network [17, 24, 31].

[1]www.flickr.com visited February 2010

More recently, information diffusion in technological networks raises interesting connections to theoretical models [19]. As news and stories become available in the real-world, they spread online forming information cascades. The capacity to collect and analyze information cascades can be useful in various domains, such as providing insight into public opinion on a variety of topics [1, 14, 30] or developing better predictive models of the spread of ideas and behaviors [2, 15, 22, 25].

Other researchers have investigated forms of cascading effects in the Flickr social graph. Singla et Weber [34] monitor the phenomenon of brand congruence in Flickr for hundreds of thousands of users and over a period of two years. Among other observations, their study suggests that two friends have a higher probability of being using the same brand, compared to two random users, suggesting the existence of social influence effects. Che et al. [6], study the distribution of photo popularity in the Flickr social graph by monitoring the favoring mechanism (users can flag photos as favorite). Their study suggests that information spreading is limited to individuals who are within close proximity of the uploaders, suggesting that social influence happens on the level of direct friends.

Characterizing the relationship between user behavior and their social environment is also the focus of other research work recently. Christakis and Fowler [16, 26] investigate cascading effects of individual behavior in human social networks. Online, Singla and Richardson [35] focus on Instant Messenger interactions and try to characterize the relationship between a person's social group and its personal behavior. They apply data mining techniques to quantify how similarity is altered based on various attributes, such as communication time with another user and more, while in [7], the authors try to quantify how social interactions affect personal interests and vice versa. A similar approach has been taken in the work of Leskovec et al. [23] where the focus was on node arrival and edge creation actions that collectively lead to macroscopic properties of networks and in [21], where authors study the diffusion of recommendations in a network by controlling a single threshold value that determines whether a user will forward a recommendation. Characterizing individual behavior in social systems has also been the topic of interest in [33], but authors do not investigate the relation of behavior and social influence. Our research complements these works in that we are interested in investigating the causality of social influence in social systems due to individual behavior, thus exploring the interplay between macroscopic and microscopic properties in the diffusion process.

3. DETECTING SOCIAL INFLUENCE

Identifying situations where social influence is present in a social system is important. In the presence of social influence, an idea, norm of behavior, or a product diffuses through the social network like an epidemic. We assume that social influence occurs in a social system when an individual's thoughts or actions are affected by one of his friends in the network.

Formally, assume a directed graph $G(V, E)$ where nodes V represent users and edges E represent friendships among users. Suppose user a adopts an innovation at time t_1. We say that user a influences user b if and only if at time t_2 when user b adopts the innovation, user a has already adopted it at an earlier time t_1, at which time a and b were already friends. We therefore assume that social influence occurs when the information of a friend adopting the innovation has the *potential* to flow to neighboring nodes in the social network.

Note that we narrow the definition of social influence to the special case where a user has at least one friend in the network that has previously adopted the innovation. We show that we can make this assumption without any loss of generality and can be generalized to more strict definitions of social influence. Our definition of social influence does not aim to be universal, but rather to help distinguish the existence of some sort of social influence from randomness in the process of adopting an innovation in a social system.

3.1 Methodology

We would like to be able to reason about whether - and if yes, in what extent - social influence occurs during the adoption of an innovation in a social system. Depending on whether they have adopted the innovation or not, nodes in the social graph $G(V, E)$ are distinguished between *active* or *inactive* respectively. We monitor the adoption of the innovation during a period $[0, T]$ and assume that activation is a binary decision, thus a node in the network that adopts the innovation remains active for the rest of the period, otherwise it remains inactive. At time $t = 0$ all nodes are inactive, while at the end of the monitoring period, $t = T$, a set A of k nodes in the network have been activated.

We now describe a method for determining if the activation of the nodes can be attributed to social influence. The idea is to test whether the activation of nodes over the monitoring period $[0, T]$ happens randomly or is affected by earlier activation of neighboring nodes in the network. If the latter is true, then we can assume that some social influence takes place, and therefore nodes that have already active neighbors have a higher probability of being activated. Note that the time of activations is important in determining social influence. Let $A = \{v_1, v_2, ..., v_k\}$ be the set of k users that are activated during the period $[0, T]$ and assume that user v_i is activated at time t_i. Since activations happen in discrete time, a total ordering of all k activation times is possible. To determine whether there is a social influence, we run the following test, called the *shuffle test*, which consists of two steps.

- **Step 1 (Original)**: We observe the adoption of the innovation in the social graph in the time period $[0, T]$. For each node activation we determine whether it can be attributed to social influence according to our definition of social influence (i.e., at least one friend is already active). Let $A_{SI} \subseteq A$ denote the set of activations that can be attributed to social influence. Then the total effect of social influence $SI_{original}$ in adopting the innovation is given by:

$$SI_{original} = \frac{A_{SI}}{A}$$

- **Step 2 (Shuffled)**: Next we create a second problem instance with the same graph G and the same set A of active nodes, by picking a random permutation π of $\{1, 2, ..., k\}$ and setting the time of activation of node v_i to $t'_i := t_{\pi(i)}$. Again we observe the adoption of the innovation in the period $[0, T]$ as in Step 1, and determine the new set $A'_{SI} \subseteq A$ of activations that can be attributed to social influence and the total effect of social influence $SI_{shuffled}$ in adopting the innovation is given by:

$$SI_{shuffled} = \frac{A'_{SI}}{A}$$

If $SI_{original} > SI_{shuffled}$ then we can conclude that some sort of social influence has been detected in the adoption of the innovation. Note that we can monitor the effect of social influence at any specific time $t \in [0, T]$ (or when a specific number of activations has occurred) by comparing the number of activations that are due to social influence as opposed to random activations by the following formula:

$$SI^t_{t\in[0,T]} = SI^t_{original} - SI^t_{shuffled}$$

Table 1: Contacts Per User

	min	max	avg	stddev	median
indegree	1	27404	89.2	270	15
outdegree	1	19542	99.9	309	19

Table 2: Contacts Per Geotagger

	min	max	avg	stddev	median
indegree	1	14664	104	277.4	28
outdegree	1	14432	104	299.2	30

Table 3: Geotagged Photos Per Geotagger

min	max	avg	stddev	median
1	92782	144	580	21

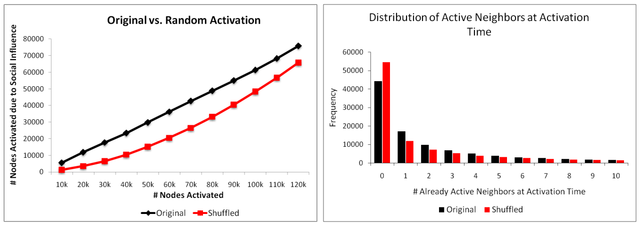
The test is based on the idea that if social influence plays an important role in adopting an innovation, then the timing of activations is not independent but rather is affected by the number of already activated neighboring nodes in the network. The idea of using timestamp permutations to distinguish influence from correlation in a social network hinges on the shuffle test suggested in Anagnostopoulos and al. [3]. However, our method does not make any assumption about a priori knowledge of the distribution of each node's influence over its neighbors (this information cannot be assumed to be known in a real system) and does not need to simulate a theoretical cascade model to determine which nodes are eventually activated. A similar randomization technique has been used by La Fond and Neville [20] to measure the gain that is due to influence and/or homophily in a social network. Their method assumes that users have attributes (e.g., age or gender) that are assumed to be known. Das et al. [8] describe sampling based methods to efficiently collect such information from a social graph. Again, our method does not make any assumption about node attributes, and as such, it is simpler, more practical and designed to efficiently be applied to other online social systems where we would like to both *detect* and *quantify* the existence of social influence.

3.2 Adoption of the Geotagging Innovation in Flickr

We apply our method of *detecting social influence* in Flickr. Flickr is a popular online social system centered around photo sharing, where users upload and share photos, comment on and tag their own photos and the photos of others, establish "friend" links and join groups of other users. Our analysis of social influence is based on the adoption of the *geotagging innovation* in Flickr.

3.2.1 Data

We crawled a large social network of the Flickr graph. To guarantee that our data includes a large number of users that have adopted the geotagging innovation, we started with a seed set of the 100 most active geotaggers and collected their contacts, and their contacts' contacts. Our final social graph consists of 525,000 nodes and 47 million directed edges. (Note that friendships are not reciprocal in Flickr.) From these users, 120,000 users have used the geotagging activity at least once (geotaggers). Table 1 gives descriptive information about the network structure for all users, and Table 2 for geotaggers only. Moreover, Table 3 provides information about the distribution of photos per geotagger. There are approximately 13 million geotagged photos in our data.



(a) *At least 1 active neighbor* (b) *Activation Histogram*

Figure 1: Original vs. Shuffled Activations

We run the shuffle test on the Flickr social graph and observe the diffusion of the geotagging innovation (i.e., node activations) for the original and the shuffled timestamps. Figure 1(a) shows that for a fixed number of activations, the number of nodes activated at a given timestamp is larger in the original timestamps than the shuffled ones. Therefore we attribute adoption of the geotagging innovation to the social influence of users. This effect is more obvious in the beginning of the diffusion process where the chance of having a random active node as a neighbor is smaller. As the process unfolds the network is filled with more and more active nodes. This increases the chance of having at least one active neighbor node and eventually causes the process to saturate. Saturation occurs when all subsequent activations can be attributed to social influence in both the original and the shuffled scenarios (i.e., upper right part of Figure 1(a)).

So far we have constrained our definition of social influence to having at least one active neighbor at the time of activation. We examine how this is affected by restricting the definition of social influence to having more active neighbors at the time of activation. To generalize on this observation, we show in Figure 1(b) the distribution of the number of already active nodes at the time that a node is activated for both the original and the shuffled timestamps. Note that in the case of the shuffled timestamps the number of nodes that have no other active neighbor at the time of activation (the initiators) is larger than that of the original timestamps. This alone means that initiators are distributed more uniformly in the graph in the case of the shuffled timestamps than in the original timestamps. This is a first indication that activations are not random. Moreover, note that for all the subsequent cases (i.e., at least 1, at least 2, at least 3, ..., at least 10 active nodes at the time of activation) the number of active neighbors is always larger in the case of the original than in the shuffled scenario. Therefore, even for more restricted definitions of social influence, there always exist an essential part of activations that may be attributed to social influence as opposed to random activations.

From this analysis we conclude that the diffusion of the geotagging innovation in Flickr (i.e., users start geotagging photos) is not random but can to a large extent be attributed to social influence of neighboring users in the Flickr social network. This observation provides evidence of a cascading process that takes place in the social graph; a process of users passing to their neighbors a signal of access to the geotagging innovation.

4. USER GEOTAGGING CREDIBILITY

We have seen evidence that users are influenced by their friends to start geotagging photos. In this Section, we develop a methodology that allows to characterize the *quality* of the geotagging behavior of users. As a surrogate of quality we define a credibility metric, based on the accuracy with which users make use of the geotagging innovation. Note that our data contains two types of geotags; either coming from GPS-enabled devices or manually assigned by a user. GPS-enabled devices give very accurate coordinates. But, to develop a user credibility metric we are more interested in the second type of geotags, which are determined by the system when a user manually places a photo on the Earth map.

The next paragraphs provide details of our methodology. In brief, it involves steps that allow to determine with high probability the location that was intented by the user (i.e., the target location) and then to measure the geodesic distance between the actual location (where the photo was placed on the earth map) and the target location. More specifically, we try to associate each user v in the network with a credibility value c_v that expresses the accuracy with which she geotags photos. Eventually, the method allows to compare a user's geotagging accuracy to the accuracy of other users and to determine who is more or less credible. As such, the method provides a primitive way to assess and argue about the *quality* of a user's expressed behavior.

Part of this work is performed by using freely available services of Yahoo! Geo Technologies [39], such as Yahoo! PlaceMaker and Yahoo! GeoPlanet.

4.1 Distinguishing Location Tags from Free-text Tags

Each photo in the dataset is associated with a set of free-text tags or tags, which are non-hierarchical keywords assigned informally to a photo by its owner. This metadata provides clues to both the context and the content of the photo and allows it to be found by browsing or searching. Tags often refer to geographical places. Throughout the paper, we refer to these tags as *location tags*.

To identify location tags, we utilize Yahoo! Placemaker, a geoparsing Web service. Provided with text, the service returns the probability that the text is a placename (e.g, the probability of the word "altitude" or "urban area" to be a placename is 0.077 and 0.055 respectively, while the probability of the word "lake echo" or "quebec city" to be a placename is 0.71 and 0.89 respectively). The Yahoo! Placemaker service takes care of the data cleansing process and considers variants of placenames in the identification process. Furthermore, the probability model (text is recognized as place with a certain probability) provides flexibility in which tags to accept as location tags, by filtering out those that are below a threshold θ.

For each photo we compute the set of its location tags by scanning the set of its tags and omitting those that are unlikely to refer to geographical places. For our needs, a threshold of 0.7 is appropriate in identifying location tags with high precision. More specifically, from an initial set of 6,756,605 unique tags, distributed over approx. 13.3M photos, we were able to identify 241,072 unique location tags. Note that this is the number of distinct location tags, so even if a location tag is very popular (e.g., "paris") it only appears once in the set. (See examples in Figure 2.)

4.2 Mapping Location Tags to Places

Location tags may be defined at various levels of granularity, for example, Country, Province, Town, etc. Yahoo! GeoPlanet uses a hierarchical model for places that provides both vertical consistency and horizontal consistency of place geography. Spatial entities in Geoplanet are identified by a unique 32-bit identifier: the *Where On Earth ID* or *WOEID*. Every named place represented by a WOEID can be mapped to a *location type* in the hierarchical model (e.g., Country, Town, Point Of Interest-POI, etc.). We would like to find a one-to-one mapping of the photo to a place, given the set of its location tags.

(a) *Gaudi's Sagrada Familia in Barcelona, Spain* (b) *Waterfalls in Edessa, Greece*

Figure 2: Distinguishing Location Tags from Free-text Tags. (a) *Tags* = {sagrada familia, gaudi, barcelona, modernist, cathedral, construction, cranes, infrared}, *Location tags* = {sagrada familia, barcelona}. (b) *Tags* = {waterfall, edessa, greece, green, pond, wall, tree, fall, balcony}, *Location Tags* = {edessa, greece}

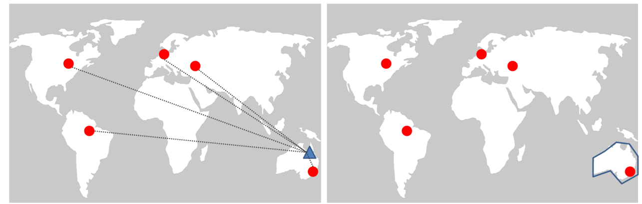
Normally a geographic name alone is not sufficient to identify a geographic place unambiguously because it may have been assigned to more than one place [9]. This happens particularly with descriptive names (for example, Newcastle, Takayama (Japanese = high mountain) and Matsushima (Japanese = pine island)), names of saints, or emigrated communities (for example, Athens). However, it is reasonable to assume that people deliberately use unambiguous placenames within local areas. Under this assumption it is possible to eliminate the problem of duplication of geographical place using two heuristics:

- **Heuristic 1: Focus the Search Near a Reference Point**: If we can focus the search of the geographical place around or close to another reference point (i.e., another place), then a location tag has a higher chance of uniquely identifying a place under this restriction (See Figure 3(a)).

- **Heuristic 2: Knowledge of Higher Levels in a Hierarchy of Places**: If the wider area in which the geographical place resides (a higher level in a hierarchy of places, such as Country, etc.) can be identified, then a location tag has a higher chance of uniquely identifying a place under this restriction (See Figure 3(b)).

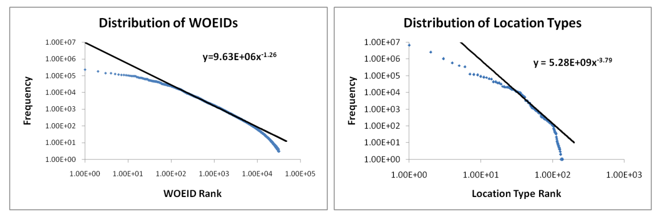
The first heuristic is always available. Each time a photo is placed on a map, latitude and longitude coordinates are assigned. These coordinates can be used to define a reference point place that guides the search and helps to disambiguate the location mention from other geographical places with the same name. The second heuristic offers further support in the disambiguation process when a photo has been assigned more than one location tag. In this case, it is expected that these can be mapped to different levels in a hierarchy of places. For example "Toronto, Canada" is mapped to (Town, Country) or "Sagrada Familia, Barcelona, Spain" is mapped to (Point of Interest, Town, Country).

Using the heuristics described above and Yahoo! GeoPlanet, we are able to efficiently disambiguate geographical names, to derive the *target location*. Each query submitted to Geoplanet, returns a



(a) *Heuristic 1* (b) *Heuristic 2*

Figure 3: Disambiguation Heuristics: Assume that red dots represent *places* on Earth with the same *location tag*. Our method efficiently disambiguates places by searching for a place that satisfies the location tag and in addition (a) is closer to a reference point (i.e., closer to the triangle on the map) or/and (b) resides in the specified boundaries of a known higher level in the hierarchy of places (i.e., Country is known)



(a) *WOEIDs* (b) *Location Types*

Figure 4: WOEID and Location Type Frequency Distributions

list of places (WOEIDs) ordered from the most likely to the least likely. Due to the well-defined semantics of the above heuristics, in almost all cases we are able to identify the most likely WOEID with a probability larger than 0.9. Thus, we are able to almost always unambiguously identify the target location by its intended WOEID.

4.3 Defining and Assessing User Credibility

Manually placing photos on a map is not an easy task and can be inaccurate at times. In our research, we focus on manually geotagged photos and try to determine the reliability of individuals by assessing the accuracy with which they place photos on the map; the closer to the target, the more reliable a user is. We measure the proximity of a user's geotagging to the target location by computing their geodesic distance d (see next paragraph). The coordinates of the target location are assumed to be known through its WOEID (all WOEIDs in our dataset are assigned a number of attributes by third-party authoritative sources, such as longitude and latitude, which we consider as the ground truth in our research). Recall that given a photo and its set of location tags we know how to unambiguously map it to its WOEID with high probability. For locations with a surface area (such as a city, as opposed to a building or statue), a wide range of latitude/longitude coordinates will fall within its boundaries. The coordinates associated with a WOEID indicate the centroid of the bounding box that encompasses the location. It is possible that the centroid of the bounding box is not the same location as the human-centric logical center of the location. Therefore, rather than comparing the user's accuracy to the coordinates of the WOEID directly, we compare the user's accuracy, to the average accuracy of all users. We assume that in the aggregate, in large numbers, users are accurate, and can capture place boundaries with a sophistication that is not represented by a centroid-based coordinate system. An example of this phenomenon is the work of Tom Taylor.[2]

4.3.1 Geodesic Distance d

Geodesic distance is the distance of two points measured along the surface of the Earth. Calculating geodesic distance is typically based on some level of abstraction, which ignores changes in elevation and other irregularities in Earth's surface. Common abstractions assume a flat, spherical, or ellipsoidal Earth. We employ an ellipsoidal approximation of the surface and compute the *ellipsoidal distance* between two points. The Ellipsoidal distance is the shortest distance between two points along the surface of an ellipsoid. It is more accurate than methods such as great-circle distance which assume a spherical Earth. We use the inverse method of Vincenty's formula[37]. This method has been widely used in geodesy because it is accurate to within 0.5 mm on the Earth ellipsoid. In our experiments we compute geodesic distances using the $WGS84$ standard Earth reference ellipsoid, used by the Global Positioning System.

4.3.2 The Wisdom of the Crowd

The geodesic distance provides a measure of proximity of the user's geotagging to the target location, but presents little information on how reliable a user is in comparison to others. Moreover, geotagging happens at different levels of location granularity. For example one user geotags at the city level and another at the country level. Thus, it is not valid to compare user credibility based on different sets of photos and different sets of intended target locations. To compensate for this limitation, we rely on a wisdom of the crowd practice. The wisdom of the crowd refers to the process of taking into account the collective opinion of a group of individuals rather than a single expert to answer a question. In our case, we would like to compare the performance of a user to the collective behavior of all users that have tried to geotag a photo that corresponds to the same target location.

Let $W = \{w_1, w_2, ..., w_k\}$ be the set of all distinct WOEIDs in our data after mapping each photo to a WOEID. Figure 4(a) plots on a log-log scale the distribution of the WOEIDs (the number of times a given WOEID has been a target location). Let a random variable D represent geodesic distances in the users' geotagging activity. For a specific WOEID w_i the random variable D takes on N real values $d_1^{w_i}, d_2^{w_i}, ..., d_N^{w_i}$, with arithmetic mean $\overline{d^{w_i}}$ and standard deviation σ^{w_i}. We compute these values ($\overline{d^{w_i}}$ and σ^{w_i}) for all WOEIDs in our data and use them to characterize the quality of the geotagging behavior of users, through hits and misses.

4.3.3 Hits and Misses

We consider geotagging to be a binary decision: a *hit* or a *miss*. We define a user's geotagging to be a hit when $d_v^{w_i}$ (the geodesic distance of user's v actual geotagging to the target place w_i) is smaller than the average distance $\overline{d^{w_i}}$ of all users geotaggings for this target place plus λ times the standard deviation σ^{w_i} of these distances. By definition, if it is not a hit, then it is considered a miss. More specifically:

$$hit : d_v^{w_i} < \overline{d^{w_i}} + \lambda \cdot \sigma^{w_i}$$

$$miss : d_v^{w_i} \geq \overline{d^{w_i}} + \lambda \cdot \sigma^{w_i}$$

where v is a user, and λ is a parameter that controls the selectivity of hits and misses in the geotagging process. A larger λ indicates

[2]http://tomtaylor.co.uk/projects/boundaries/ visited February 2010.

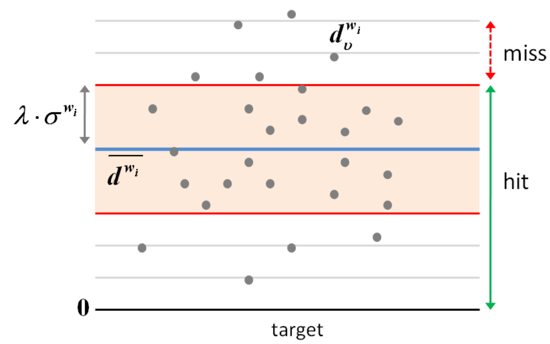


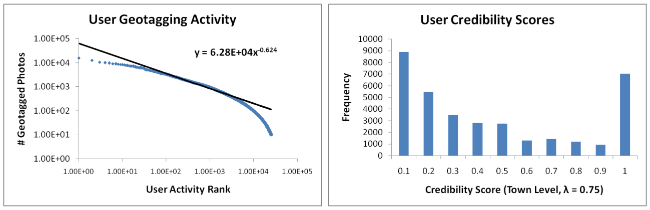
Figure 5: Illustrative example of the way a *Hit* (or *Miss*) is defined for a specific target location w_i following a wisdom of the crowd practice and provided the distances of all users' geotaggings to the target location

that a hit is easier (i.e., larger geodesic distance from the target is allowed) and vice versa (See Figure 5).

Recall that our initial objective was to associate each user v with a single credibility value c_v. However, individuals may tag photos at different levels in the place hierarchy (different *location types*) and a WOEID can be mapped to any location type. Figure 4(b) plots in a log-log scale the distribution of the location types in our data (that is, the number of times a target place was of that location type). Determining the semantics of a user credibility in a single value that aggregates information from all location types (Country, Town, POI, etc.) is challenging (as semantics of the hierarchy of places could be violated). Instead, we prefer to maintain *a vector representation* of a user's credibility, let $C_v = \{c_v^{l_1}, c_v^{l_2}, ..., c_v^{l_n}\}$, where each vector dimension represents the user credibility at a specific location type. We define the credibility value of a user v at a specific location type l_i (e.g., Town) to be the number of hits $h_v^{l_i}$ for a given location type, divided by the total number of geotagged photos by this user of that location type $t_v^{l_i}$. Note here, that we are allowed to aggregate averages over all WOEIDs of the same location type, despite the fact that some places are more popular than other. This is because our definition of hits and misses incorporates these semantics (i.e., despite how popular a place is, in order to get a hit you must be at least as accurate as many other people that tried to geotag photos for the same place). Our methodology aims to provide common ground in assessing the quality of the geotagging behavior. Designing measures of behavior quality in a generic way is challenging as it largely depends on the application of interest and, as such, is out of the scope of this paper.

5. INFLUENCE VS. BEHAVIOR

Understanding the way in which an individual's choices depend on what other individuals do requires analysis that can be performed at two conceptually different levels of network resolution. A global one, in which we observe network effects in aggregate, and a local one, in which we observe how individuals are influenced by their network neighbors [10]. In Section 3, we detected the occurrence of social influence in the adoption of the geotagging innovation in Flickr by monitoring aggregate information of user actions on a macroscopic scale. In Section 4, we focused on individuals and monitored user behavior on a microscopic scale. In this section we try to find points where these two different levels of resolution converge and investigate the causality of social influence.



(a) *User Geotagging Activity* (b) *Credibility Scores Histogram*

Figure 6: User Geotagging Behavior

Many of our interactions with the rest of the world happen at a local, rather than a global, level - we often don't care as much about the full population's decisions as about the decisions made by friends and colleagues. To this end, we bring the analysis closer to the detailed network level and look at how individuals are influenced by their particular network neighbors. A particular individual behavior, such as geotagging, can be communicated to the network in different ways. In our research we focus on (a) intensity and on (b) credibility of geotagging. These properties are distinct as intensity of geotagging characterizes the *quantity* of the expressed behavior (i.e., how many times a user expressed the behavior), while credibility of geotagging characterizes the *quality* of the expressed behavior (i.e., how well a user expressed the behavior). Our research examines which of the two primitive types of the expressed behavior (related to quantity vs. related to quality) has more effect in the influence that a user exerts in her social network. Associating aggregate observations with individual behaviors improves our understanding of how individual's choices depend on what others do and eventually how ideas, behaviors and innovations diffuse in a population.

5.1 Data

As mentioned earlier, different users may be more or less accurate for different location types. We consider a user's credibility for each location type independently. In the following experiments we focus on the credibility of users at the *town* level. Limiting our analysis to a single location type provides more clear semantics when comparing users. We pick *towns* because they are the most popular location type in our data. In addition, there is a very large variation in user credibility scores at the town level.

To produce a more reliable data set, we reject users who only occasionally used the geotagging technology or who were extremely inaccurate in identifying the places associated with their photos. Thus we eliminated users who had fewer than ten photos geotagged as well as users that had a credibility score of zero. The final set consisted of approximately 25,000 users that have geotagged more than 5 million photos. Figure 6(a) plots in a log-log scale the distribution of user geotagging activity. The x-axis represents the rank of each user, taking values from 1 to approximately 25,000, ordered by activity from the most to the least active.

Recall that λ is a parameter that controls the selectivity of hits and misses in the geotagging process, and setting it correctly is dependent on the particular application. Since we are investigating in a general setting, with no specific application in mind, we seek a setting that provides variability in the credibility scores. We experimented with different settings of λ, and found that making λ large (i.e., 1.0 and over) causes most users to be highly credible, while making λ small (i.e., 0.5 and below) is too strict and no users are credible. For values between 0.5 and 1.0, most users are assigned intermediate credibility scores. For our experiments we fix lambda to 0.75 as it provides sufficient variability in the credibility scores of users. Figure 6(b) shows a histogram of the credibility scores for $\lambda = 0.75$. Still, the majority of users have low credibility scores. The peak at credibility of one may be because all of the user's photos were from one location, or because the location was sufficiently large that the user could drop it in the appropriate place on the map. It may also be the case of users that tend to geotag photos of their hometown, thus being very accurate, or the images were taken with GPS-enabled cameras.

5.2 Experiment Sketch

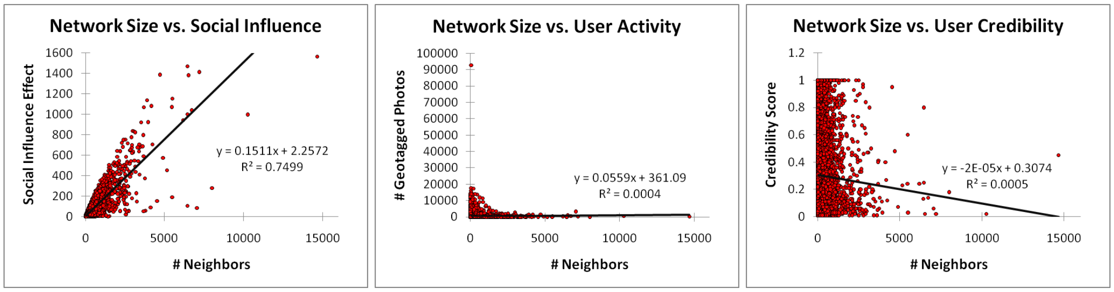
We design an experiment that measures how influential is a user u to its social network, in adopting the geotagging innovation. The semantics hinge on the social influence experiment discussed in Section 3, but differ in focus. Here the focus is not on determining if a user's activation was due to a neighbor's influence, but rather the opposite, to evaluate how influential a user is in her network, i.e. her *social influence effect*. Formally, we say that user a has the potential to influence user b, if and only if user a adopted the geotagging technology at time t_1, and user b adopted it at time t_2, and a and b became friends at time t_3, where $t_1 < t_2$ *and* $t_3 < t_2$. That is, a user a has the potential to influence a user b, if a was activated before b and they became friends at least before b was activated. Therefore a user is said to influence a neighbor if he has the *potential* to pass to his neighbor a signal of access to the geotagging innovation.

Given these semantics, it is possible to algorithmically assess the potential social influence of a user (or a set of users V) to the network over a time period. Algorithm 1 gives the details of a recursive computational procedure for evaluating social influence in a finite number of steps. It takes as input a set of users V, a social graph G and the activation times of all users A in the monitoring period $[0, T]$, and returns the number of users that have potentially been influenced by the set of users V.

In our experiments we do not assume that users have the same amount of time to influence others. In fact, earlier adopters have larger amount of time. It is important to note though that we monitor the adoption of a real technology from its launch (time 0 in our experiments), and therefore provide to all users "equal" opportunity to be early adopters (through a form of external influence; reading about it, testing the innovation, etc.). Could late adopters complain that they didn't have the chance or the time to influence others? Yes, but this is always the case in a real setting (e.g., political influence, fads, diffusion of innovation, etc.) and we would like to be able to capture this effect. Algorithm 1 is the basic tool to assess and compare the *social influence effect* of varying sets of users in our experiments.

Effect of Neighborhood Size: Before investigating the effect of individual behaviors to the social influence, it is important to examine whether the number of neighbors that an individual has, plays an important role in determining her social influence. To determine this effect we explore the correlation between the network size of an individual (i.e., number of neighbors in our context) and her social influence effect.

Figure 7 shows scatter plots and the regression lines for the correlation tests we run (Pearson's correlation was used as a measure of correlation). Our analysis shows that there is direct and high correlation between the neighborhood size of an individual and the social influence she exerts to her social network (Figure 7(a)). This is in alignment with [4], where authors report that adoption rates quicken as the number of friends adopting increases and this effect varies with the connectivity of a particular user. Due to the large effect that the neighborhood size has to an individual's social in-



(a) *Network Size vs. Social Influence* (b) *Network Size vs. User Activity* (c) *Network Size vs. User Credibility*

Figure 7: Network Size Effect

Algorithm 1 Social Influence Effect Estimation

```
procedure EVALSOCIALINFLUENCE(V, G, A)
    V: Set of users
    G: Social Graph (from, to, friendship_create_time)
    A: Activation Times (user, activation_time)
    I: Set of Influenced Users

    I ← ∅
    for all v ∈ V do
        t1 = getActivationTime(v, A)
        N^v = getFriends(v, G)
        for all n ∈ N^v do
            t2 = getActivationTime(n, A)
            t3 = getFriendshipCreateTime(v, n, G)
            if t1 < t2 && t3 < t2 then
                I ← I ∪ n
            end if
        end for
    end for
    return |I|
end procedure
```

fuence, for the rest of the experiments we had to normalize on the neighborhood size parameter. The normalization allows to focus on the individual behavior observations. Otherwise, these observations could simply be a function of larger or smaller neighborhood. From now on, any time we refer to the social influence effect of a user, we refer to a proportion that represents the number of influenced neighbors over the total number of neighbors.

It's also important to note that as shown in Figure 7(b) and Figure 7(c), there is very weak to negligible correlation between an individual's neighborhood size and any of the two behaviors under investigatation. In other words, the number of the neighbors has no effect in the expressed behaviors of an individual.

5.2.1 *Social Influence vs. User Activity*

We hypothesize that users who are more active are more influential. This stands to reason because users who are more active upload more photos, and provide geotags for more photos, they are "experts" of a sort, in the social system of user-generated content. To test the hypothesis we run experiments that measure the relationship between user activity (as determined by the number of photos a user has geotagged) and influence in the Flickr social network.

From the set of 25,000 users, we first sort the users based on their activity (number of geotagged photos), from the most active to the least active. Then, we define 5 user groups, each consisting of 5,000 users, based on their activity level. The first group has the most active users and the last one the least active users (Figure 8(a)). For each activity level we compute the potential social influence in the network using Algorithm 1. We report the median value of the normalized social influence effect, as the distribution of this property is very skewed.

Assuming the increased visibility they have in their network such as through notification mechanisms in Flickr, active users, because they generate more content, are viewed by more people. In fact, we see in Figure 8(b) that more active users can be as much as 23% more influential in the social network, in terms of encouraging people to adopt the innovation of geotagging.

5.2.2 *Social Influence vs. User Credibility*

The measure of credibility based on geodesic distance is intuitive: a person is more accurate as a geotagger if she consistently places her photos very close to the true intended location. However credibility may relate not only to how accurate a person is, but how accurate they are perceived to be by others. Thus there is an interplay between influence and accuracy. We hypothesize that more credible users are more influential. To test the hypothesis we run experiments that measure the relationship between accuracy and influence in the Flickr social network.

From the set of 25,000 users, we first sort the users based on their credibility score, from the most credible to the least credible. Then, we define 5 user groups, each consisting of 5,000 users, based on their credibility level. The first group has the most credible users and the last one the least credible users (Figure 9(a)). For each user group we compute the potential social influence in the network using Algorithm 1. We report the median value of the normalized social influence effect, as the distribution of this property is very skewed.

One would expect that users with higher credibility scores - that is, users who are more accurate - would be more influential, assuming that people adopt a technology because they have seen it done well. But in fact, as we show in Figure 9(b), this is not the case. It appears that the reasons people become good geotaggers do not relate to social influence. That is, you may adopt the technology when you see another person try it, but whether they are good at it or not does not factor into your decision to adopt the technology. We further discuss the consequences of this result in the next paragraph.

Hidden Social Influence: This is an unexpected finding, as credibility of a user consists an essential individual quality that appears to be currently hidden in the social network. This is an important observation, since it suggests that current design of online social systems may refrain users from exploiting the full potential of their social influence. In practice, it lessens or cancels the *social influence* of a user in her environment. This is not happening intentionally, but it's rather a consequence of stale design. Interfaces to electronic systems have traditionally been designed as single-user systems. The existence of other users and their activities have been implicitly assumed to be an attribute of the system that should be hidden from end-users. Similar design approach has been adopted for accessing information on the web.

This observation raises a more general concern. While *quantity* of behavior is usually clearly communicated in social systems, *quality* of behavior is systematically not (e.g., due to lack of feedback mechanisms related to quality). As users interact in online systems they expose or develop different qualities of individual behavior. This becomes more evident in modern online social systems, where behaviors may or (may not) be revealed by the system to their social peers. For example, the quality of user-generated content in Twitter[3] or in Facebook[4] varies drastically from being of high quality to being abuse or spam. Quality of user behavior consists an important part of a user's social value, which might translate to power of influencing her environment or decision making processes, but largely remains hidden. As such, our research suggests that users should have the option to reveal certain qualities of their behavior to their social network. Acknowledging the social value of users offers a design opportunity; in designing social systems, it is not only necessary to see other users, but to also clearly communicate what behaviors are disclosed and how they are formed - a design that entails a balance of visibility, awareness of others and their qualities.

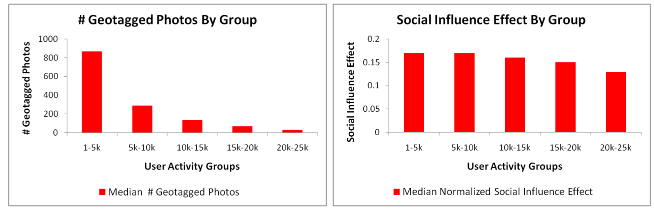
5.3 Tests of Significance

Thoughout our study we based our observations on sample populations. But, how dependent our observations on these samples are? We performed statistical significance tests for the two behaviors in our study (i.e., user activity and user credibility) to assess evidence in favor of or against our claims. The observations in the samples (i.e., groups of users) are all real numbers that describe the median value of the normalized social influence in each group of individuals. The values of the property we observe typically has skewed distributions. We actually run normality tests, based on the Shapiro-Wilk method, to show that the samples in each group do not follow a normal distribution. Thus, to compare k-samples (i.e., the 5 different groups in each behavior) we employ a method that does not assume a normal population of samples, such as the Kruskal-Wallis test, for a selected significance level $alpha = 0.05$. This test is a non-parametric method for testing equality of population medians among groups. Intuitively, it is identical to a one-way analysis of variance with the data replaced by their ranks. In the case that the null hypothesis is rejected by the Kruskal-Wallis test, we perform post-hoc analysis to identify which of the groups differ. We use the bonferroni adjustments post-hoc test to identify these pairs. For the various cases in our study, we formally define the following hypotheses:

H_0: *The samples are not significantly different*
H_a: *The samples do not come from the same population*

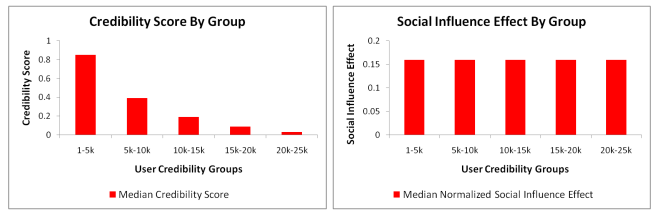
For the case of Figure 8(b) (Social Influence vs. User Activity), the Kruskal-Wallis test showed that the samples do not come from the same population, thus rejecting the null hypothesis H_0. The post-hoc analysis revealed that all groups were significantly different with each other. This supported our claim that more active users are more influential in the social network.

For the case of Figure 9(b) (Social Influence vs. User Credibility), the Kruskal-Wallis test showed that the samples come from the same population, thus accepting the null hypothesis H_0. This supported our claim that being more credible does not necessarily make a user more influential in the social network.



(a) *User Activity By Group* (b) *Social Influence vs. User Activity*

Figure 8: Social Influence vs. User Activity



(a) *User Credibility By Group* (b) *Social Influence vs. User Credibility*

Figure 9: Social Influence vs. User Credibility

6. CONCLUSIONS

We developed a method that detects social influence in a social system. The method is similar to the method proposed in [3], as they both hinge on the idea of shuffling the timestamps of social actions in a network. Despite its technical soundness and completeness, their method makes assumptions of prior knowledge of the distribution of a node's influence over its neighbors and of a theoretical cascade model that when is simulated determines which nodes are eventually activated. Our method is simpler and is designed to both detect and quantify a user-defined social influence in the context of a real social system. We applied our method in Flickr to demonstrate that adoption of the geotagging innovation can to a large extent be attributed to social influence between users. This observation provides evidence of a cascading process taking place in the social graph; a process of users passing to their neighbors a signal of access to the geotagging innovation. The proposed method is generic and can be useful for providing decision support for designing viral marketing campaigns in social systems.

We performed a series of experimental and observational studies on rich data coming from a large social system in order to investigate causality, and in particular to draw a conclusion on the effect of changes of user behavior in determining how influential they are in their network. In particular, we studied the geotagging behavior of individuals that collectively leads to macroscopic properties of social influence and contagion. Our main objective was to investigate potential points of convergence between the macroscopic and microscopic view of the social system. This is significant because

[3] www.twitter.com
[4] www.facebook.com

it allows us to argue about higher levels of granularity by observing individual behavior.

We investigated the role of the neighborhood size in an individual's behavior. Our findings suggest that individuals that have many neighbors have more potential to influence their friends in their social network. Then, we investigated two types of user behavior, frequency of using the geotagging innovation and accuracy in placing images on a map, and tried to assess the effect of the assumed behavior on its social influence in the network. Note that these two behaviors are interesting, as the former characterizes the quantity of the expressed behavior, while the latter characterizes the quality of the expressed behavior.

Our findings suggest that the first is a potential *reinforcing activity* such that when it occurs, the probability of a friend in the network adopting the behavior increases. This is an important observation that relates to the concept of *reinforcement* in behavior sciences [5, 36]. On the other hand, we found that the credibility of a user has only a small effect in how influential she is in her network. This raises concerns for the design of social systems, since user credibility, by construction, is an important indicator of a user's quality.

To the best of our knowledge, our work is the first that experimentally unveils that there are qualities of individual behavior that remain hidden, a practice that lowers the social influence of individuals in their network. As an outcome of our research, we propose the conception of mechanisms that inform users about their friends' activities and qualities, hoping that such a social feedback would eventually help users to regain their hidden social value and encourage other people in the network to expose or develop high quality behaviors. Similar systems are seen in [28, 29] where users become aware of similar people's surfing behavior and in Yahoo! Answers[5] where users are encouraged to answer accurately within a reward system based on a token economy.

There are many open issues left for future research. For instance, in our experiments we assume that social influence occurs only through internal stimuli, realized in Flickr through social feedback mechanisms. As such we do not take into consideration any external stimuli or source of communication. However, social systems co-exist (for example Flickr, Blogosphere, our family, our classmates and more) and communication between social systems is possible [27]. Formal verification of social influence occurrence is tricky since it is susceptible to the assumptions made about the social system. Being able to argue about causality of social influence on the whole is a challenging task that requires the design of controlled user experiments.

7. REFERENCES

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD*, 2005.

[2] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *WSDM*, 2008.

[3] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD*, 2008.

[4] E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In *EC*, 2009.

[5] D. E. Blackman. *Operant Conditioning: An Experimental Analysis of Behaviour*. Routledge Kegan & Paul, 1974.

[6] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW*, 2009.

[7] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD*, 2008.

[5]http://answers.yahoo.com visited February 2010

[8] G. Das, N. Koudas, M. Papagelis, and S. Puttaswamy. Efficient sampling of information in social networks. In *SSM*, 2008.

[9] M. Doerr and M. Papagelis. A method for estimating the precision of placename matching. *IEEE Trans. on Knowl. and Data Eng.*, 19(8), 2007.

[10] D. Easley and J. Kleinberg. Cascading behavior in networks. In *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[11] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[12] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Mark. Lett.*, 12(3), 2001.

[13] M. Granovetter. Threshold models of collective behavior. *The American Journal of Sociology*, 83(6), 1978.

[14] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD*, 2005.

[15] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW*, 2004.

[16] N. A. C. James H. Fowler. Cooperative behavior cascades in human social networks. *PNAS*, 107(12), 2010.

[17] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[18] J. Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. In N. Nisan, T. Roughgarden, Éva Tardos, and V. Vazirani, editors, *Algorithmic Game Theory*. Cambridge University Press, 2007.

[19] J. Kleinberg. The convergence of social and technological networks. *Commun. ACM*, 51(11), 2008.

[20] T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW*, 2010.

[21] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), 2007.

[22] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, 2009.

[23] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD*, 2008.

[24] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007.

[25] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *SDM*, 2007.

[26] J. H. F. Nicholas A. Christakis. The spread of obesity in a large social network over 32 years. *N Engl J Med.*, 357(4), 2007.

[27] L. Niklas. The world society as a social system. *International Journal of General Systems*, 8(3), 1982.

[28] A. Papagelis, M. Papagelis, and C. Zaroliagis. Enabling social navigation on the web. In *WI-IAT*, 2008.

[29] A. Papagelis, M. Papagelis, and C. Zaroliagis. Iclone: towards online social navigation. In *HT*, 2008.

[30] M. Papagelis, N. Bansal, and N. Koudas. Information cascades in the blogosphere: A look behind the curtain. In *ICWSM*, 2009.

[31] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, 2002.

[32] E. M. Rogers. *Diffusion of innovations*. Free Press, 5th edition, 2003.

[33] E. Santos-Neto, D. Condon, N. Andrade, A. Iamnitchi, and M. Ripeanu. Individual and social behavior in tagging systems. In *HT*, 2009.

[34] A. Singla and I. Weber. Camera brand congruence in the flickr social graph. In *WSDM*, 2009.

[35] P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *WWW*, 2008.

[36] B. F. Skinner. *The Behavior of Organisms. An Experimental Analysis*. Appleton-Century-Crofts, 1938.

[37] T. Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, 22(176), 1975.

[38] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge Univ Pr, 1994.

[39] Yahoo! Geoplanet. *http://developer.yahoo.com/geo/*, 2009.