Human Mobility, Social Ties, and Link Prediction

Dashun Wang[1,2] Dino Pedreschi[1,3] Chaoming Song[1,2] Fosca Giannotti[1,4] Albert-László Barabási[1,2,5]

[1]CCNR, Dept. of Physics and Computer Science, Northeastern University, Boston, MA 02115, USA
[2]CCSB, Dana-Farber Cancer Institute, Harvard University, Boston, MA 02115, USA
[3]KDD Lab, Dipartimento di Informatica, Università di Pisa, 56127 Pisa, Italy
[4]KDD Lab, ISTI-CNR, Istituto di Scienza e Tecnologie dell'Informazione del C.N.R., 56124 Pisa, Italy
[5]Dept. of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
{dashunwang, dino.pedreschi, chaoming.song, fosca.giannotti, barabasi}@gmail.com

ABSTRACT

Our understanding of how individual mobility patterns shape and impact the social network is limited, but is essential for a deeper understanding of network dynamics and evolution. This question is largely unexplored, partly due to the difficulty in obtaining large-scale society-wide data that simultaneously capture the dynamical information on individual movements and social interactions. Here we address this challenge for the first time by tracking the trajectories and communication records of 6 Million mobile phone users. We find that the similarity between two individuals' movements strongly correlates with their proximity in the social network. We further investigate how the predictive power hidden in such correlations can be exploited to address a challenging problem: which new links will develop in a social network. We show that mobility measures alone yield surprising predictive power, comparable to traditional network-based measures. Furthermore, the prediction accuracy can be significantly improved by learning a supervised classifier based on combined mobility and network measures. We believe our findings on the interplay of mobility patterns and social ties offer new perspectives on not only link prediction but also network dynamics.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms

Measurement, Performance

Keywords

Human Mobility, Link Prediction, Social Network

KDD'11, August 21–24, 2011, San Diego, California, USA.

1. INTRODUCTION

Social networks have attracted particular interest in recent years, largely because of their critical role in various applications [11, 5]. Despite the recent explosion of research in this area, the bulk of work has focused on the social space only, leaving an important question of to what extent individual mobility patterns shape and impact the social network, largely unexplored. Indeed, social links are often driven by spatial proximity, from job- and family-imposed programs to joint involvement in various social activities [28]. These shared social foci and face-to-face interactions, represented as overlap in individuals' trajectories, are expected to have significant impact on the structure of social networks, from the maintenance of long-lasting friendships to the formation of new links.

Our knowledge of the interplay between individual mobility and social network is limited, partly due to the difficulty in collecting large-scale data that record, simultaneously, dynamical traces of individual movements and social interactions. This situation is changing rapidly, however, thanks to the pervasive use of mobile phones. Indeed, the records of mobile communications collected by telecommunication carriers provide extensive proxy of individual trajectories and social relationships, by keeping track of each phone call between any two parties and the localization in space and time of the party that initiates the call. The high penetration of mobile phones implies that such data captures a large fraction of the population of an entire country. The availability of these massive CDRs (Call Detail Record) has made possible, for instance, the empirical validation in a large-scale setting of traditional social network hypotheses such as Granovetter's strength of weak ties [27], the development of a first generation of realistic models of human mobility [14, 29] and its predictability [30]. Indeed, despite the inhomogeneous spatial resolution (the uneven reception area of mobile phone towers) and sampling rates (the timing of calls), the large volume of CDR data allows us to reconstruct many salient aspects of individual daily routines, such as the most frequently visited locations, and the time and periodicity of such visits. Therefore, these data serve as an unprecedented social microscope helping us scrutinize the mobility patterns together with social structure and the intensity of social interactions.

In this work, we follow the trajectories and communication patterns of approximately 6 Million users over three months, by using CDR data from an anonymous country, aiming to measure for any pair of users u and v:

- *How similar is the movement of u and v.* For this purpose, we introduce a series of *co-location* measures quantifying the similarity between their movement routines, prompting us to call them the *mobile homophily* between u and v.
- *How connected are u and v in the social network.* For this purpose, we adopt several well-established measures of network proximity, based on the common neighbors or the structure of the paths connecting u and v in the who-calls-whom network.
- *How intense is the interaction between u and v.* For this purpose we use the number of calls between u and v as a measure of the strength of their tie.

Our analysis offers empirical evidence that these three facets, co-location, network proximity and tie strength, are positively correlated with each other. In particular, we find that the higher the mobile homophily of u and v, the higher the chance that u and v are strongly connected in the social network, and that they have intense direct interactions. These findings uncover how the social network, made of numerous explicit who-calls-whom ties, is embedded into an underlying mobility network, made with the implicit ties dictated by the mobile homophily.

The emergence of such surprising three-fold correlation hints that it is conceivable, to some extent, to predict one of the three aspects given the other two. Indeed, we demonstrate in this study how the predictive power hidden in these correlations can be exploited to identify new ties that are about to develop in a social network. Specifically, we study the influence of co-location and mobile homophily in link prediction problems, asking: what is the performance of mobility-based measures in predicting new links, and can we predict more precisely whether two users u and v (that did not call each other in the past) will call each other in the future, by combining the measurements of their network proximity *and* mobile homophily? Our key findings are summarized as follows:

- The mobility measures on their own carry remarkably high predictive power, comparable to that of network proximity measures.
- By combining both mobility and network measures, we manage to significantly boost the predictive performance in supervised classification, detecting interesting niches of new links very precisely. For example, by considering a subset of potential links (pairs of users) with high network proximity and mobile homophily, we are able to learn a decision-tree classifier with a precision of 73.5% and a recall of 66.1% on the positive class. In other words, only approximately one fourth of the predicted new links were false positives, and only one third of the actual new links were missed by the predictor.

To the best of our knowledge, this work presents the first assessment of the extent individuals' daily routines as a determinant of social ties, from empirical analysis to prediction models. With recent proliferating advances on human mobility and social networks, we believe our findings are of fundamental importance in our understanding of human behavior, provide significant insights towards not only link prediction problems but also the evolution and dynamics of networks, and could potentially impact a wide array of areas, from privacy implications to urban planning and epidemic prevention.

2. MOBILE PHONE DATA

Currently the most comprehensive data that contains simultaneously both human mobility and social interactions across a large segment of the population is collected by mobile phone companies. Indeed, mobile phones are carried by their owners during their daily routines. As mobile carriers record for billing purposes the closest mobile tower each time the user uses his phone, the data capture in detail individual movements. With almost 100% penetration of mobile phones in industrial countries, the mobile phone network is the most comprehensive proxy of a large-scale social network currently in existence. We exploit in this study a massive CDR dataset of approximately 6 Million users, which, to the best of our knowledge, is the largest dataset analyzed to date containing both human trajectories and social interactions. We focused on 50k individuals selected as the most active users (identical to those that were studied in [30]), following not only their trajectories but also their communication records during 14 successive weeks in 2007[1].

The resulting dataset contains around 90M communication records among the individuals, and over 10k distinct locations covering a radius of more than 1000 km. Each record, for our purposes, is represented as 4-tuple $\langle x, y, t, l\rangle$, where user x is the caller, user y is the callee, t is the time of the call, and l is the location of the tower that routed the call. The temporal granularity used in this study is the hour, justified by the findings in [14, 30, 29]. Let V denotes the set of users. For each user $x \in V$, the total number of calls initiated by x is denoted as $n(x)$. For x's i-th communication, where $1 \leq i \leq n(x)$, the time stamp, location, and the contacted user are denoted as $T_i(x)$, $L_i(x)$ and $N_i(x)$, respectively. Given a time interval between t_0 and t_1, the set of communications between pairs of users occurred within the interval is denoted as $E[t_0, t_1] \equiv \{(x, y)|x, y \in V, \exists i, 1 \leq i \leq n(x), t_0 \leq T_i(x) < t_1, N_i(x) = y\}$. In other words, we add an edge (x, y) if there has been at least one communication between x and y in the interval. Therefore, $G[t_0, t_1] \equiv \{V, E[t_0, t_1]\}$ is the resulting social network within the time interval.

To prepare for the link prediction experiments, we further separate our data into 2 parts: first 9 weeks for constructing the old network and the rest 5 weeks for the new network. For each link $e \in E$, we classify it according to its time stamp $t(e)$. $E_t \equiv \{e|e \in E, t \leq t(e) < t + 1\}$ is defined as the set of edges of the resulting network after aggregating the communications in the t-th week. The "past" and "future" sets are therefore denoted as $E_{old} = \bigcup_{t=1}^{9} E_t$ and $E_{new} = \bigcup_{t=10}^{14} E_t$. In our study, we focus on nodes in the largest connected component $G_{old} = \{V_{old}, E_{old}\}$, where we observe in total $|V_{old}| = 34,034$ users and $|E_{old}| = 51,951$ links.

3. NETWORK PROXIMITY

General approaches in link prediction tasks have been focused on defining effective network based "proximity" measures, so that two nodes that are close enough on the graph

[1]Unfortunately, we cannot make this dataset available due to NDA.

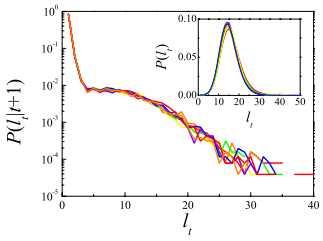


Figure 1: The probability density function $P(l_t|t+1)$ that a new link has chemical distance l in previous week. Inset: the probability density function of chemical $P(l_t)$ for different weeks.

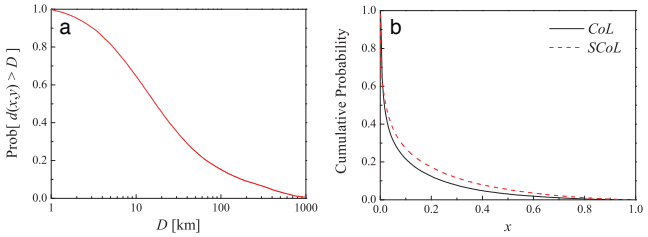


Figure 2: a) The probability two users i and j have distance $d(i,j) > D$. b) The probability two users i and j have Co-Location $CoL(i,j)$ (solid) and Spatial Co-Location $SCoL(i;j)$ (dashed) greater than x.

but not yet connected may have a better likelihood of becoming connected in the future. As the main focus of the paper is to explore the predictive power of mobility compared and combined with topological predictors, we selected four representative quantities which have been proven to perform reasonably well in previous studies (for more details of the quantities and their performance on citation networks, see [21].)

- *Common neighbors.* The number of neighbors that nodes x and y have in common. That is, $CN(x,y) \equiv |\Gamma(x) \cap \Gamma(y)|$, where $\Gamma(x) \equiv \{y|y \in V, (x,y) \in E\}$ is the set of neighbors of x.

- *Adamic-Adar* [1]. A refinement of $CN(x,y)$ by weighting common neighbors based on their degrees, instead of simple counting. Therefore the contribution from hubs to common neighbors is penalized by the inverse logarithm of their degree.
$AA(x,y) \equiv \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$.

- *Jaccard's coefficient.* Defined as the size of the intersection of the neighbors of two nodes, $\Gamma(x)$ and $\Gamma(y)$, divided by the size of their union, characterizing the similarity between their sets of neighbors.
$J(x,y) \equiv |\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)|$.

- *Katz* [17]. Summation over all possible paths from x to y with exponential damping by length to weight short paths more heavily. $K(x,y) \equiv \sum_{l=1}^{\infty} \beta^l \cdot |paths^l_{x,y}|$, where $paths^l_{x,y}$ is the set of all paths with length l from x to y (damping factor β is typically set to 0.05.)

Most network proximity measures are related to the chemical distance on the graph, under the natural assumption that new links are more likely to occur between nodes that are within a small distance on the graph. The *chemical distance* $l(x,y|E)$ is defined as the length of the shortest path between two nodes x and y. $l(x,y|E) = 1$ implies that nodes x and y are connected, or $(x,y) \in E$. The role of chemical distance on tie formation can be tested directly by measuring the probability $P(l_t|t+1)$ for a new link $e \equiv (x,y) \in E_{t+1}$ to have a chemical distance l_t measured at previous week t. That is, $P(l_t|t+1) \equiv |\{e|e \equiv (x,y) \in E_{t+1}, l(x,y|E_t) = l_t\}| / |E_{t+1}|$. This distribution is shown in Fig. 1, different colors indicating different time windows t. We find, first of all, $P(l_t|t+1)$ is stable over different weeks (1 through 14), indicating that the aggregation process we adopted to construct the network is robust, and that $P(l_t|t+1)$ is largely independent wrt the time windows. Second, $P(l_t|t+1)$ decays rapidly as l_t increases, consistent with previous studies [20] on other data sets. This implies that the majority of new links are between nodes within two hops from each other, i.e., nodes with common neighbors. Third, the Poisson distribution of the chemical distance for arbitrary pairs (inset of Fig. 1) suggests that the most probable distance for two users to form a link at random is around 12, while it is only 2 for pairs that do form new links.

4. MOBILE HOMOPHILY

Similar to the graph-based approaches, a natural strategy to predict new links by leveraging mobility information is to look for quantities that capture some degree of closeness in physical space between two individuals. Indeed, people who share high degree of overlap in their trajectories are expected to have a better likelihood of forming new links [28]. Therefore, we explored a series of quantities aiming to define the similarity in mobility patterns of two individuals.

- *Distance.* Let
$$ML(x) \equiv \text{argmax}_{l \in Loc} PV(x,l)$$
be the most likely location of user x, where Loc is the set of all locations (cell phone towers), and
$$PV(x,l) \equiv \sum_{i=1}^{n(x)} \delta\left(l, L_i(x)\right) / n(x)$$
is the probability that user x visits location l[2]. We define $d(x,y) \equiv \text{dist}(ML(x), ML(y))$ as the distance between two users x and y, representing the physical distance between their most frequented locations.

- *Spatial Co-Location Rate.* The probability that users x and y visit at the same location, not necessarily at the same time. Assuming that the probability of visits for any two users are independent, we define:
$$SCoL(x,y) \equiv \sum_{l \in Loc} PV(x,l) \times PV(y,l)$$

- *Spatial Cosine Similarity.* The cosine similarity of user x and y's trajectories, capturing how similar their visitation frequencies are, assigned by the cosine of the angle between the two vectors of number of visits at each location for x and y.
$$SCos(x,y) \equiv \sum_{l \in Loc} \frac{PV(x,l) \times PV(y,l)}{\|PV(x,l)\| \times \|PV(y,l)\|}$$

[2]Here $\delta(a,b) = 1$ if $a = b$, 0 otherwise.

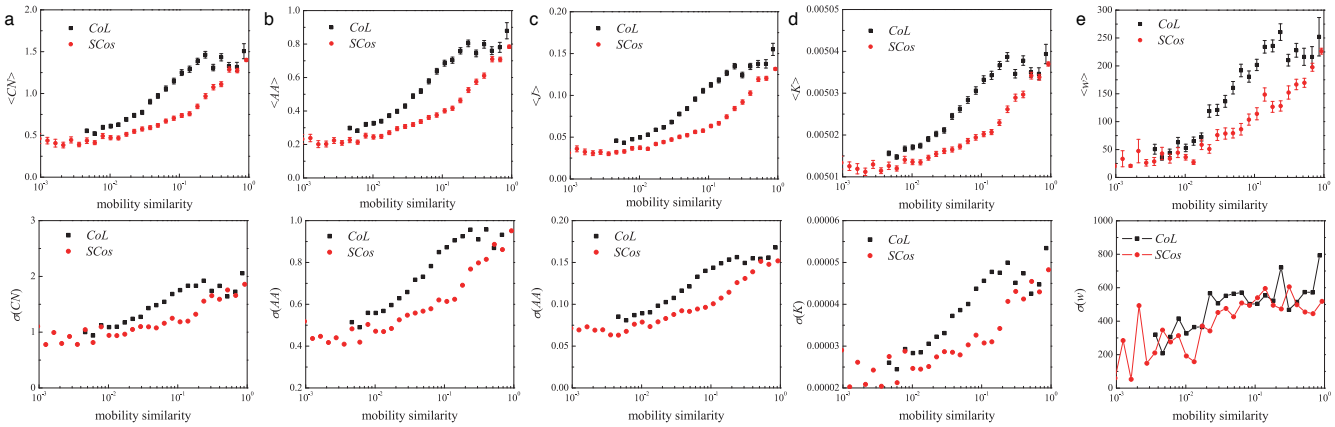


Figure 3: Correlations between mobility measures (CoL and $SCos$) and a) Common Neighbor, b) Adamic Adar, c) Jaccard Coefficient, d) Katz, and e) link weight. The upper panels show the mean values, whereas the lower panels show the standard deviations

- *Weighted Spatial Cosine Similarity.* The *tf-idf* version of cosine similarity of the visitation frequencies of users x and y, where the contribution of each location l is inversely proportional to the (log of) its overall population in l. Coherent with the *tf-idf* idea in information retrieval, this measure promotes co-location in low-density areas, while penalizes co-location in populated places.

- *Co-Location Rate.* The probability for users x and y to appear at the same location during the same time frame (hour):

$$CoL \equiv \frac{\sum_{i=1}^{n(x)} \sum_{j=1}^{n(y)} \Theta\left(\Delta T - |T_i(x) - T_j(y)|\right) \delta\left(L_i(x), L_j(y)\right)}{\sum_{i=1}^{n(x)} \sum_{j=1}^{n(y)} \Theta\left(\Delta T - |T_i(x) - T_j(y)|\right)}$$

  where $\Theta(x)$ is the Heaviside step function, and ΔT is set to 1 hour. This quantity takes into account the simultaneous visits of two users at the same location, i.e., both spatial and temporal proximity, normalized by the number of times they are both observed at the same time frame.

- *Weighted Co-Location Rate.* The *tf-idf* version of CoL, i.e., the probability for two users x and y to co-locate during the same hour, normalized by the (log of) population density of the co-location at that hour.

- *Extra-role Co-Location Rate.* The probability for two users x and y to co-locate in the same hour at night or during weekends. As shown in [10], close proximity of two individuals during off-hours may serve as a powerful predictor for symmetric friendships.

The quantities listed above either aim at measuring the geographical closeness or the degree of trajectory overlap of two individuals, characterizing their mobile homophily. It should be noted that it is not obvious whether the spatiotemporal co-location measures, e.g., CoL, would yield better estimates of the probability of face-to-face interactions than spatial only measures, e.g., $SCoL$. Indeed, on one hand, CoL quantifies the co-presence of two users in the same place around the same moments, corresponding to a high likelihood of meeting face-to-face. Yet there are circumstances where two users do co-locate but are not captured by the data if any one of them did not place any phone calls. And this latter case is captured to some extent by $SCoL$, as the necessary condition for two individuals to meet is the spatial overlap of their trajectories.

We now explore the distributions of the various measures over the linked pairs of individuals $(x, y) \in E_{old}$. In Fig. 2a we show the complementary cumulative distribution function (CCDF) of geographical distances $d(x, y)$. We find that $d(x, y)$ follows a fat-tailed distribution, consistent with previous studies [19, 18, 22], meaning that while most friends live close to each other, there are also friends who are far apart. The CCDF plots of CoL and $SCoL$ are shown in Fig. 2b as solid and dashed line, respectively. $SCoL$ measures the probability for two users to appear at the same location, capturing, spatially, the degree of trajectory overlapping. CoL quantifies the probability of appearing at the same place around the same time, characterizing the spatio-temporal overlap of trajectories. We find that "friends" typically do co-locate, in that most pairs $(x, y) \in E_{old}$ exhibit non-zero spatial or spatio-temporal overlap in their trajectories, and such overlap decays very fast.

5. CORRELATION BETWEEN MOBILE HOMOPHILY AND NETWORK PROXIMITY

We explore a series of connections between similarity in individual mobility patterns and social proximity in the call graph, by measuring the correlation between the proposed mobility and network quantities, using again the edges in G_{old}. We also consider the strength of the ties in the network, quantified by the number of calls placed between any two users (during the first 9 weeks of our observation period.) In Fig. 3, we plot the mean values and the standard deviations of Common neighbors, Adamic-Adar, Jaccard's coefficient, Katz, and the strength of social ties for different values of Co-Location and Spatial Cosine Similarity, discretized by logarithmic binning. We find that the quantities that characterize the proximity in the social graph systematically correlate with mobility measures. The more similar two users' mobility patterns are, the higher is the chance

Table 1: Pearson Coefficients

	CoL	$SCos$	J	CN	AA	K	w	dML
CoL	1	0.76286	0.25359	0.19618	0.2251	0.18952	0.14521	-0.17894
$SCos$	0.76286	1	0.30789	0.25657	0.28679	0.24933	0.14402	-0.24938
J	0.25359	0.30789	1	0.82384	0.88147	0.81108	0.11348	-0.10136
CN	0.19618	0.25657	0.82384	1	0.94437	0.99939	0.05989	-0.098562
AA	0.2251	0.28679	0.88147	0.94437	1	0.93806	0.086881	-0.10126
K	0.18952	0.24933	0.81108	0.99939	0.93806	1	0.053842	-0.095631
w	0.14521	0.14402	0.11348	0.05989	0.086881	0.053842	1	-0.029339
dML	-0.17894	-0.24938	-0.10136	-0.098562	-0.10126	-0.095631	-0.029339	1

that they have close proximity in the social network, as well as the higher is the intensity of their interactions. Furthermore, Fig. 4 demonstrates that the geographical distance between two individuals decays logarithmically with mobility measures. We omit the plots where the network proximity measures and the tie strength are on the x-axis, due to space limitations, but we observe a qualitatively similar trend in all cases. The Pearson coefficients of each pairs of variables are reported in Table 1. It is interesting to observe that tie strength, although conceived as a network measure, is more strongly correlated with mobile homophily than with network proximity measures.

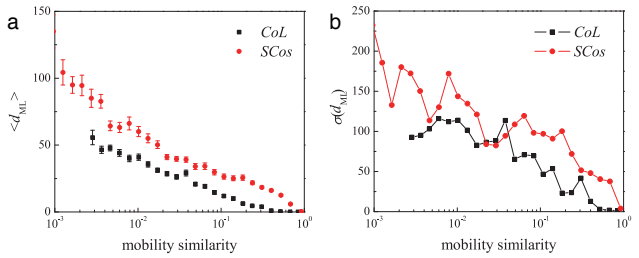


Figure 4: Correlations between mobility measures (CoL and $SCos$) and distance between two individuals. (a) mean values and (b) standard deviations.

Taken together, our results indicate that mobile homophily, network proximity and tie strength strongly correlate with each other. This fact implies that mobile homophily is a viable alternate candidate to predict network structures, and motivates the investigation of a novel approach to link prediction that takes into account both mobility and network measures. Moreover, we find that the standard deviation for the correlation plots are not small, hinting that there are extra degrees of freedom which allow us to further improve our predictive power by using supervised classification methods combining the mobility and network dimensions together.

6. LINK PREDICTION

6.1 Design of the link prediction experiment

We now study the link prediction problem in the context of our mobile social network. Link prediction is a classification problem, aimed at detecting, among all possible pairs of users that did not call each other in the past, those that will communicate in the future. We define a *potential link* any pair of users (u, v) such that $(u, v) \notin E_{old}$, i.e., users u and v did not call each other from week 1 through 9, and a *new link* any potential link (u, v) such that $(u, v) \in E_{new}$, i.e., users u and v did not call each other from week 1 through week 9, but did call each other (at least once) from week 10 through week 14. Finally, we define a *missing link* any potential link which is not a new link, i.e., a pair of users that did not call each other in the entire period from week 1 through week 14. For any potential link (u, v), let $NL(u, v)$ be a binary variable with value 1 if (u, v) is a new link, and 0 if (u, v) is a missing link.

In this setting, link prediction is formalized as a binary classification problem over the set of all potential links, where the class label is specified by the NL variable, and the predictive variables are the network and mobility quantities introduced in Sec. 3 and 4, measured over the first period from week 1 through week 9. According to this formulation, we aim to predict whether a potential link becomes a new link in the "future" based on the observation of its "past" network connectedness and mobile homophily.

Our dataset consists of $n = 34,034$ users and $m = 51,951$ links, resulting in $(n(n-1)/2) - m = 579,087,610$ potential links. Yet the actual new links are only $12,484$ – about 2 new links every 10^5 potential links! The significant number of potential links creates obvious computational challenges, both in terms of memory and time. Moreover, the huge disproportion between new links and missing links implies an extreme unbalance between the positive and negative class, which makes the classification task prohibitive. To cope with both difficulties, we followed two complementary strategies for selecting subsets of potential links: *i*) *progressive sampling*: we consider increasingly large samples of missing links, up to some manageable size, and *ii*) *links with common neighbors*: we concentrate on the interesting case of pairs of nodes that are two hops away in the network, i.e., nodes with common neighbors, and consider the entire population of potential links between such nodes. We report below the results obtained in our link prediction analysis in both cases.

Another dimension of our study is the kind of classification used. Adhering to the machine learning terminology [23], we consider both *unsupervised* and *supervised* link prediction:

- The unsupervised method, originally proposed in [21], consists in ranking the set of potential links using one of the available network or mobility quantities, and then classifying as new links the k top-ranked potential links, where k is the expected number of new links (as

measured in the dataset.) The rest are classified as missing links.

- The supervised method consists in learning a classifier, e.g., a decision tree, using a training set of new links and missing links, and then classifying each pair as a new or missing link according to the class assigned by the learned classifier.

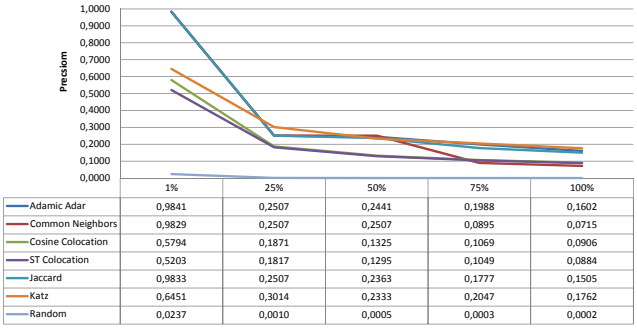
Different unsupervised classifiers are obtained by considering the various network and mobility measures, and different supervised classifiers are obtained by considering different combinations of the same quantities as predictive variables. We systematically constructed the complete repertoire of classifiers, based either on network quantities, or mobility quantities, or the combination of the two. We then compared their quality and predictive power. To this extent, we put particular attention on the metric used to assess a classifier, given that simple accuracy (over either the training or test set) is a misleading measure for classifiers learned over highly unbalanced datasets. Indeed, recall that in our case the trivial classifier that labels each potential link as missing has a 99.998% accuracy. The real challenge in link prediction is achieving high *precision* and *recall* over positive cases (new links), defined in terms of the confusion matrix of a classifier (see Table 2): *precision* $= \frac{TP}{TP+FP}$, and *recall* $= \frac{TP}{TP+FN}$. Traditionally, precision and recall are combined into their harmonic mean, the F-measure. However, we put more emphasis on precision, as the most challenging task is to classify some potential links as new links with high probability, even at the price of a non negligible number of false negatives. We also use lift and gain charts to compare the precision of the various classifiers over the percentiles of the examined test cases.

6.2 Progressive sampling of missing links

In our first set of experiments we created various unsupervised and supervised classifiers over the complete dataset of positive cases, i.e., 12,484 new links, augmented with up to 51M negative cases of missing links. We assess the precision achieved by each classifier when used with all 12,484 new links and increasing fractions of missing links, i.e., to 1%, 25%, 50%, 75% and 100% of the total 51M missing links sampled. Figure 5 summarizes our findings for unsupervised classifiers. For each network/mobility quantity Q and each dataset with increasing samples of missing links, we rank the potential links in the dataset for decreasing values of Q, and the top ranked 12,484 links are predicted as new links. Each line in Fig. 5 describes how the precision for different quantities decays with the size of missing links. On the positive side, all unsupervised classifiers are significantly better than random guessing, and the decay of their precision tends to stabilize. Nevertheless, as these 51M links are only about 10% of the total missing links, we conclude that all quantities exhibit modest predictive power. The most surprising finding is that the co-location measures have a comparable precision to network measures: slightly worse than best network predictors (Katz, Adamic-Adar), but better than Common Neighbors. Moreover, mobility measures have a slower decay than network measures over increasing negative sample size. The observation that the two classes of measures have approximately similar predictive power offer further evidence that social connectedness is strongly correlated with mobile homophily.

	predicted class = 0	predicted class = 1
actual class = 0	TN (true neg.)	FP (false pos.)
actual class = 1	FN (false neg.)	TP (true pos.)

Table 2: Confusion matrix of a binary classifier



	1%	25%	50%	75%	100%
Adamic Adar	0,9841	0,2507	0,2441	0,1988	0,1602
Common Neighbors	0,9829	0,2507	0,2507	0,0895	0,0715
Cosine Colocation	0,5794	0,1871	0,1325	0,1069	0,0906
ST Colocation	0,5203	0,1817	0,1295	0,1049	0,0884
Jaccard	0,9833	0,2507	0,2363	0,1777	0,1505
Katz	0,6451	0,3014	0,2333	0,2047	0,1762
Random	0,0237	0,0010	0,0005	0,0003	0,0002

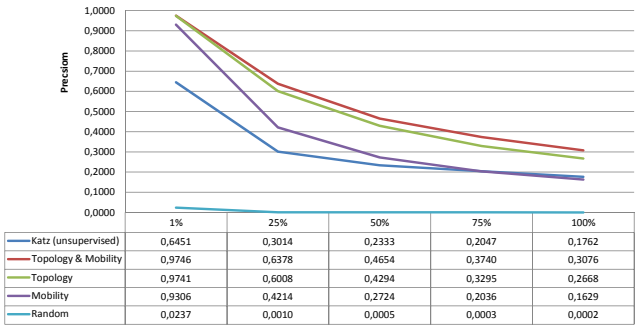
Figure 5: Precision of unsupervised classifiers over increasing fractions of missing links (1%, 25%, 50%, 75% and 100% of the total 51M missing links sampled). Ranking is obtained using the various network and mobility measure. Precision refers to the fraction of new links among the top-ranked 12,484 potential links. The precision of the random classifier is shown as baseline.

Figure 6 illustrates the supervised case: we consider the best classifiers obtained using network and mobility measures, both in isolation and combined together. Once again, we consider negative samples of increasing size, up to 51M missing links, and measure the decay of precision as in the unsupervised case. We considered a vast repertoire of classification algorithms (decision trees, random forests, SVM, logistic regression) under diverse parameter settings, and report in the chart the most robust classifiers, evaluated with cross validation, with strongest evidence against overfitting. In the chart we also compare the precision of the supervised methods (evaluated on an independent test set) with that of the best unsupervised predictor (Katz). We observe that the precision of the supervised classifiers is about double of their unsupervised counterpart, and mobility measures once again achieve comparable predictive powers to the traditional network measures. The best precision, around 30% in the 51M case, is obtained using the network and mobility measures combined together. Therefore, using network measures in combination with co-location measures yields a sensible improvement. Indeed, the probability of correctly predicting a new link is 1500+ times larger than random guessing.

6.3 Potential links with common neighbors

To get better insight, we concentrate on the nodes that are two hops away from each other in E_{old}, i.e., all potential links (u, v) of mobile users in our complete network such that u and v have at least one common neighbor during the first two months. The motivations behind this approach are two-fold. First, most new links that do form belong to this category (Fig. 1), and we hope to boost our prediction models by focusing on this most promising set of links. Second, by focusing on these links, the total number of potential links becomes computationally manageable, which enables us to assess the asymptotic behavior of prediction accuracy.

There are 266,750 potential links in this case, of which 3,130 (1.17%) formed a new link. Note that, different from



	1%	25%	50%	75%	100%
Katz (unsupervised)	0,6451	0,3014	0,2333	0,2047	0,1762
Topology & Mobility	0,9746	0,6378	0,4654	0,3740	0,3076
Topology	0,9741	0,6008	0,4294	0,3295	0,2668
Mobility	0,9306	0,4214	0,2724	0,2036	0,1629
Random	0,0237	0,0010	0,0005	0,0003	0,0002

Figure 6: Precision of the best supervised classifiers found over increasing fractions of missing links (1%, 25%, 50%, 75% and 100% of the total 51M missing links sampled), using only network measures, only mobility measures, and combination of both. Precision of best unsupervised classifier (K) and random classifier is shown as baseline.

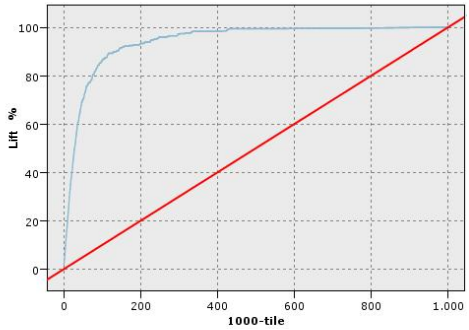


Figure 7: Lift chart of the best decision tree found in the dataset of potential links with common neighbors; the x-axis represents the percentiles of the potential links in the test set ranked by decreasing probability of being new links, as specified by the learned classifier. A point (x, y) in the blue curve represents the fact that $y\%$ of the actual new links are found when considering the top-ranked $x\%$ potential links predicted as positive. The red straight line is the lift of the random classifier. In our classifier, more than 85% new links are found considering only the 10% most probable positive potential links.

the previous case, we now consider the entire population of missing links. We study the precision of the unsupervised and supervised methods in this case. In the unsupervised case, the precision for the different measures is computed by considering the fraction of new links in the top-ranked 3,130 cases in the list ordered by the precision of each measure in descending order:

Measure	Precision
Katz	9.1%
Adamic-Adar	7.8%
Spatial Cosine Similarity	5.6%
Weighted Spatial Cosine Similarity	5.6%
Extra-role Co-Location Rate	5.1%
Weighted Co-Location Rate	5.1%
Common Neighbors	5.1%
Co-Location Rate	5.0%
Jaccard	3.0%

As we now have a complete set of negative cases, we corroborate our findings in Sec. 6.2 that mobility measures indeed yield remarkably high predictive power in the unsupervised setting, comparable to network measures in the link prediction literature. Furthermore, various mobility measures have very similar performance, indicating these measures all adequately capture the similarity in mobility patterns.

In the supervised case, after systematic, yet heuristic, exploration of a large space of classification methods with different parameters, we construct a decision-tree using Quinlan's C4.5 classification over the combined network and mobility measures, with cross validation to control over-fitting, applied to the subset of potential links with common neighbors under the further constraint $AA > 0.5$ and $SCoL > 0.7$. Our tree has the following confusion matrix over an independent test set (1 = new link), implying a precision of 73.5% and a recall of 66.1%.

	pred. class = 0	pred. class = 1
actual class = 0	6,627	82
actual class = 1	117	228

Both precision and recall are one order of magnitude larger than all previous figures. The lift chart (Fig. 7) for this classifier shows how, e.g., 86.4% of new links are found by considering only the top 10% positive cases, as ranked by the classifier in descending order of their probability of being new links. Interestingly, we find that the classifier obtained with the procedure discussed above, but using network measures only, has precision 36.2% and recall 6.1%, suggesting that the combination of topology and mobility measures is crucial to achieve high precision and recall. In other words, learning a supervised classifier based on combined network and mobility measures significantly boosts the precision and recall of predicted new links. The price to pay is that we need to focus on a niche of promising potential links with high AA and $SCos$ coefficients, concentrating on a relatively small number of candidates, yet for those we gain a very high probability of guessing the correct new links. While stressing the use of specific classification techniques, e.g., ad-hoc link prediction methods optimized for highly-unbalanced data, such as HPLP [23, 6], to achieve better precision is beyond our goals here, it is indeed an interesting open question for future research.

7. RELATED WORK

In this section, we review three categories of related work: studies on human mobility patterns, link prediction in social networks, and interplay between physical space and network structure.

7.1 Human Mobility

In the past few years, the availability of large-scale datasets, such as mobile-phone records and global-positioning-system (GPS) data, has offered researchers from various disciplines access to detailed patterns of human behavior, greatly enhancing our understanding of human mobility.

From statistical physics perspective, significant efforts have been made to understand the patterns of human mobility. Brockmann et al. [4] tested human movements using half a million dollar bills, finding that the dispersal of bills is best

modeled by continuous-time random walk (CTRW) models. González et al. [14] then showed that each individual is characterized by a time-independent travel distance and a significant probability to re-visit previous locations, by using mobile phone data of 100,000 individuals. Song et al. [29] then proposed a statistically self-consistent microscopic model for individual mobility. Researchers have also found individuals' daily routines are highly predictable, by using principal component analysis [9] and measuring mobility entropy [30].

From data mining perspective, there have been a number of studies mining frequent patterns on human movements. General approaches are based on frequent patterns and association rules, and build predictive models for future locations. To name a few, Morzy used a modified version of Apriori [25] and Prefixspan [26] algorithms to generate association rules. Jeung et al. [16] developed a hybrid approach by combining predefined motion functions with the movement patterns of the object, extracted by a modified version of the Apriori algorithm. Yavas et al. [32] predicted user movements in a mobile computing system. Furthermore, Giannotti et al. [12, 13] developed trajectory pattern mining, and applied it to predict the next location at a certain level of accuracy by using GPS data [24].

7.2 Link prediction in social networks

Link prediction has attracted much interest in recent years after the seminal work of Liben-Nowell and Kleinberg [21]. It is a significant challenge in machine learning due to the inherent extreme disproportion of positive and negative cases. Existing approaches have focused on defining various proximity measures on network topology, to serve as predictors of new links in both supervised [2, 31, 23, 15] and unsupervised [21] frameworks. Most of the empirical analyses are based on co-authorship networks, and the domain-dependent features developed in certain studies (see, e.g., [2]) are tailored to this particular data set. The supervised high-performance link prediction method HPLP in [23, 15] has also been applied to a large phone dataset, using only network proximity measures.

The fundamental difference of our study from this literature is that we focus on the impact of human mobility, an intrinsic property of human behavior, on link prediction. Indeed, we have designed a broad range of mobile homophily measures and explored their power in predicting new links. Our research is orthogonal to the above line of research, in the sense that any general link prediction method can be used in combination with our mobility features, e.g., the machine learning techniques for extremely unbalanced classes.

7.3 Interplay between physical space and network structure

Although it is in general difficult to obtain data that contain simultaneously the geographical and network information, there have been a few interesting attempts to assess the interplay between the two. For example, there is empirical evidence [19, 18, 22] showing that the probability of forming a social tie decays with distance as a power law. Based on this fact, Backstrom, et al. [3] introduced an algorithm that predicts the location of an individual. A few recent studies focused either on small populations of volunteers, whose whereabouts and social ties were monitored at fine detail using ad-hoc smart-phone applications [10] and location-sharing services [8], or on large but specific online communities such as Flickr [7]. Although none of these data could provide a society-wide picture of either social interactions or individuals' daily routines, these studies indeed indicate that the strong correlation between physical space and network structures emerges in many diverse settings.

8. CONCLUSIONS AND FUTURE WORK

Recent advances on human mobility and social networks have turned the fundamental question, to what extent do individual mobility patterns shape and impact the social network, into a crucial missing chapter in our understanding of human behavior. In this work, by following daily trajectories and communication records of 6 Million mobile phone subscribers, we address this problem for the first time, through both empirical analysis and predictive models. We find the similarity between individuals' movements, their social connectedness and the strength of interactions between them are strongly correlated with each other. Human mobility could indeed serve as a good predictor for the formation of new links, yielding comparable predictive power to traditional network-based measures. Furthermore, by combining both mobility and network measures, we show that the prediction accuracy can be significantly improved in supervised learning.

We believe our findings on the interplay of mobility patterns and social ties offer new perspectives on not only link prediction but also network dynamics. At the same time, they also have important privacy implications. Indeed, the surprising power of mobility patterns in predicting social ties indicates potential information leakage from individuals' movements to their friendship relations, posing a new challenge in privacy protection. Furthermore, we believe our results could impact a wide array of phenomena driven by human movements and social networks, from urban planning to epidemic prevention.

The results presented in this paper also open up many interesting directions for future research. The first is to search for improvement in link prediction tasks by judiciously mixing mobility and network measures. For example, we find that adding co-location measures into Adamic-Adar could yield a precision of 9.6% in unsupervised classification, overtaking any traditional measures listed in the paper. While exhaustively searching for such quantities is beyond our goals here, further work in this direction would be very important. The second direction is to study this problem in a more systematic manner by using more users and different datasets from multiple countries. Indeed, in this paper, we mostly focused on 50k active users from one single country. This approach inevitably introduced some sampling bias, impacts of which are not yet fully-understood. It is therefore important to adapt current frameworks to more users in the dataset and similar datasets from different countries. Another interesting direction is to look at the inverse problem with respect to this work. Indeed, upon uncovering the strong correlations between mobility similarity and social connectedness and predicting links based on mobility patterns, the question thereafter is can we gain more insights about individuals' whereabouts by leveraging our knowledge of their social ties and activity patterns? In sum, the increasing availability of mobile phone data and the emergence of location-based social networking websites has the power to revolutionize our understanding of the in-

terplay between mobility and social networks, making this field particularly fallow for new results.

Acknowledgement. The authors wish to thank members of CCNR in Boston and KDD Lab in Pisa for many useful discussions, and anonymous reviewers for many insightful suggestions. This work was supported by Google Research Award; the *Mobility, Data Mining and Privacy* (MODAP) EU FET OPEN Coord. Action; the NSCTA sponsored by the U.S. ARL under Agreement Number W911NF-09-2-0053; the James S. McDonnell Foundation 21st Century Initiative in Studying Complex Systems; the U.S. ONR under Agreement Number N000141010968; the DTRA awards WMD BRBAA07-J-2-0035 and BRBAA08-Per4-C-2-0033; the NSF within the Information Technology Research (DMR-0426737), and IIS-0513650 programs.

9. REFERENCES

[1] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM: Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

[3] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW*, pages 61–70, 2010.

[4] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[5] G. Caldarelli. *Scale-free networks: complex webs in nature and technology*. Oxford University Press, 2007.

[6] D. A. Cieslak and N. V. Chawla. Learning decision trees for unbalanced data. In *ECML/PKDD*, pages 241–256, 2008.

[7] D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436, 2010.

[8] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Ubicomp*, pages 119–128, New York, NY, USA, 2010. ACM.

[9] N. Eagle and A. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[10] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274, 2009.

[11] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[12] F. Giannotti, M. Nanni, and D. Pedreschi. Efficient mining of temporally annotated sequences. In *SDM*, 2006.

[13] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *KDD*, pages 330–339, 2007.

[14] M. González, C. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[15] Z. Huang, X. Li, and H. Chen. Link prediction approach to collaborative filtering. In *JCDL*, pages 141–142. ACM, 2005.

[16] H. Jeung, Q. Liu, H. T. Shen, and X. Zhou. A hybrid prediction model for moving objects. In *ICDE*, pages 70–79, 2008.

[17] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[18] G. Krings, F. Calabrese, C. Ratti, and V. Blondel. Urban gravity: a model for inter-city telecommunication flows. *J. Stat. Mech.-Theor. Exp.*, page L07003, 2009.

[19] R. Lambiotte, V. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

[20] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD*, pages 462–470, 2008.

[21] D. Liben-Nowell and J. M. Kleinberg. The link prediction problem for social networks. In *CIKM*, pages 556–559, 2003.

[22] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33):11623, 2005.

[23] R. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *KDD*, pages 243–252, 2010.

[24] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *KDD*, pages 637–646, 2009.

[25] M. Morzy. Prediction of moving object location based on frequent trajectories. In *ISCIS*, pages 583–592, 2006.

[26] M. Morzy. Mining frequent trajectories of moving objects for location prediction. In *MLDM*, pages 667–680, 2007.

[27] J. P. 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(18):7332–7336, 2007.

[28] M. Rivera, S. Soderstrom, and B. Uzzi. Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms. *Annual Review of Sociology*, 36:91–115, 2010.

[29] C. Song, T. Koren, P. Wang, and A.-L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 2010.

[30] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of predictability in human mobility. *Science*, 327(5968):1018, 2010.

[31] C. Wang, V. Satuluri, and S. Parthasarathy. Local probabilistic models for link prediction. In *ICDM*, pages 322–331, 2007.

[32] G. Yavas, D. Katsaros, Ö. Ulusoy, and Y. Manolopoulos. A data mining approach for location prediction in mobile environments. *Data Knowl. Eng.*, 54(2):121–146, 2005.