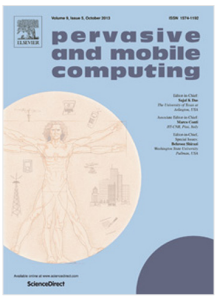
Accepted Manuscript



A novel mathematical framework for similarity-based opportunistic social networks

Mai ElSherief, Babak Alipour, Mimonah Al Qathrady, Tamer ElBatt, Ahmed Zahran, Ahmed Helmy

PII: S1574-1192(16)30354-6
DOI: http://dx.doi.org/10.1016/j.pmcj.2017.08.004
Reference: PMCJ 884

To appear in: *Pervasive and Mobile Computing*

Please cite this article as: M. ElSherief, B. Alipour, M.A. Qathrady, T. ElBatt, A. Zahran, A. Helmy, A novel mathematical framework for similarity-based opportunistic social networks, *Pervasive and Mobile Computing* (2017), http://dx.doi.org/10.1016/j.pmcj.2017.08.004

A Novel Mathematical Framework for Similarity-based Opportunistic Social Networks☆

Mai ElSherief*, Babak Alipour‡, Mimonah Al Qathrady‡, Tamer ElBatt†◇, Ahmed Zahran**◇, Ahmed Helmy‡

*Department of Computer Science, University of California, Santa Barbara (USA)
†Wireless Intelligent Networks Center (WINC), Nile University (Egypt)
**Department of Computer Science, University College Cork (Ireland)
◇Faculty of Engineering, Cairo University (Egypt)
‡Computer and Information Science and Engineering Department, University of Florida, Gainesville (USA)

Abstract

In this paper we study social networks as an enabling technology for new applications and services leveraging, largely unutilized, opportunistic mobile encounters. More specifically, we quantify mobile user similarity and introduce a novel mathematical framework, grounded in information theory, to characterize fundamental limits and quantify the performance of sample knowledge sharing strategies. First, we introduce generalized, non-temporal and temporal profile structures, beyond geographic location, as a probability mass function. Second, we examine classic and information-theoretic similarity metrics using data in the public domain. A noticeable finding is that temporal metrics give lower similarity indices on the average (i.e., conservative) compared to non-temporal metrics, due to leveraging the wealth of information in the temporal dimension. Third, we introduce a novel mathematical framework that establishes fundamental limits for knowledge sharing among similar opportunistic users. Finally, we show numerical results quantifying the cumulative knowledge gain over time and its upper bound, the knowledge gain limit, using public smartphone data for the user behavior and mobility traces, in the case of fixed as well as mobile scenarios. The presented results provide valuable insights highlighting the key role of the introduced information-theoretic framework in motivating future research along this ripe research direction, studying diverse scenarios as well as novel knowledge sharing strategies.

Keywords: social networks, opportunistic, profiles, similarity, modeling, user traces, numerical results

1. Introduction

Recent studies by the International Telecommunication Union (ITU), e.g., [1], point out that mobile phone coverage is now nearly ubiquitous, with an estimated 95% of the global population about seven billion people living in an area covered by at least a basic 2G cellular network. In

☆This work was supported in part by a Google Faculty Research Award to Nile University.

2016, more than 75% of the population in Europe and the Americas has access to mobile broadband subscriptions whereas Africa has about 30% penetration, yet, is steadily growing [1]. These global trends, complemented by a variety of wireless technologies and applications, has inspired innovative networking regimes ranging from personal and social [2, 3] to professional. However, formally characterizing and leveraging the inherent social structure exhibited by mobile users persist as major challenges hindering optimized resource allocation and new services. Recent studies have investigated bridging the divide between the social structure of users and wireless networking [4]. Prior studies in social sciences, e.g., Homophily [Lazarsfeld and Merton (1954)], demonstrate that people tend to have similarities with those in close proximity. In those communities of interest, users typically establish trust, communicate, and interact [5]. Hence, smartphones have strong potential to enhance the mobile user experience with the aid of personalized applications, e.g., location-aware services, targeted advertisement in addition to recommendation systems [6] and social networks among others.

The development of similarity-based, opportunistic social networking applications would typically involve the design of three core components, namely mobile user profiles, similarity assessment, and knowledge sharing if users are deemed similar. User profiles capture behavioral patterns relevant to the application of interest. The similarity assessment component judges, quantitatively, the similarity between the profiles of mobile users in proximity. Once two users are deemed similar, they may share knowledge and tips using policies that may depend on the service type and user preferences. For instance, two shoppers coming in proximity, in the same store (e.g., kids wear), would exchange their "anonymized" profiles to assess similarity. If similar, the smartphone application exchanges tips about stores ratings, special offers, and other relevant information. Despite the fact that establishing trust in opportunistic settings [7] and profile anonymization are key components of the envisioned system, they are complementary to this work and are important subjects for future research. In this paper, we assume trust is established among all users and focus on introducing the new mathematical framework instead.

Our prime contribution in this work is a new information-theoretic framework for opportunistic social networks established on the basis of user similarity. The major contributions of this paper can be summarized as follows:

- Generalize the mobile user profiles, beyond the user geographic location, to a probability distribution function incorporating multiple facets and study non-temporal and temporal models.
- Unveil valuable insights about legacy and new temporal and non-temporal similarity metrics, using datasets in the public domain [8]. Moreover, we draw attention to the merits of the Hellinger distance, from information theory, to assess similarity between probability distribution user profiles.

- Introduce a new temporal similarity metric, using matrix vectorization, to incorporate the important dynamics in the temporal dimension that affect similarity, yet, with lightweight computations.
- Formally define the novel concepts of *Knowledge Gain* per user, and its upper bound, *Knowledge Limit* for opportunistic device-to-device (D2D) social networks.
- Establish fundamental limits using information theoretic results and provide important insights for different topologies, knowledge sharing policies and mobility patterns and validate our findings using public domain user behavior and mobility datasets.

The remaining of this paper is organized as follows. Section 2 motivates the work and Section 3 surveys related work in the open literature. In Section 4, we study mobile user similarity with emphasis on probability distribution profiles, using classic and novel metrics. In Section 5, we shift our attention to the novel information-theoretic framework to characterize fundamental limits on performance (Knowledge Limit) and quantify the Knowledge Gain of sample policies. We present performance results based on real user mobility traces [9, 10] augmented with behavior traces from another data set [8]. Finally, we draw conclusions and provide directions for future work in Section 6.

2. Motivation

The wide spread of powerful smartphones renders them coupled to the users, keeping a treasure of valuable user behavior data, e.g., mobility, circles of interaction, applications, etc., inferring information about the mobile user's interests. This data has not only imposed research challenges but also provided new directions to enrich user experiences [11]. Examples of crowdsourcing applications that leverage user real-time mobility are Waze and Google Maps which provide road traffic congestion alerts and route alternatives.

Motivated by the strong link between users' behavior and their smartphone usage, we seek to address the following open question; *Can we leverage the knowledge and prior experiences of people with similar interests whom we encounter daily?* In order to answer this question, we propose a framework for an envisioned type of applications, called *opportunistic recommendation systems* (ORS). An example of ORS is introduced in [12], whereby mobile devices exchanges are bound by homophily. ORS enable users to extend their regular everyday "conversational" recommendations, from people they know and meet to "cyber" swaps with similar mobile users opportunistically encountered and, even further, to never encountered users, through "knowledge forwarding" discussed later.

Finally, we argue that opportunistic social networks, established based on user similarity, could give rise to a variety of new network services, e.g., establishing trust, targeted marketing, friend

recommendations, and location-aware services. In addition, ORS could promote a variety of new smartphone applications serving large public events, e.g., fair grounds, sports arenas, etc.

As a proof of concept, we implemented an Android mobile application for opportunistic recommendation systems coined O'BTW (Oh By The Way) [12] whereby mobile devices can anonymously exchange ratings, knowledge and recommendations with other "similar" users and happen to meet opportunistically. A "conference-setting" O'BTW use case was demonstrated at ACM MobiSys 2013 as shown in Figure 1. Simplified user profiles were constructed on the spot, through a simple GUI, based on conference participation over the past three years. Assuming trust is established, O'BTW first creates a Bluetooth connection between smartphones in proximity. Afterwards, it tests pair-wise, temporal profile similarity using a newly proposed metric, called Vectorized Cosine, to be discussed in Section 4. If found similar, they exchange stored recommendations representing the user's knowledge.

Figure 1: A knowledge sharing proof-of-concept Android mobile application coined "Oh By The Way" (O'BTW).

3. Related Work

There has been growing interest in mobile social networks over the past few years. In [13], a review is given for mobile social networks, highlighting proposed architectures, social properties and key research challenges. Prior work on opportunistic mobile social networks can be roughly aligned along two major thrusts. First, the routing problem has received attention, e.g., [14, 15]. In [14], the authors propose a home-aware community model towards a distributed optimal Community-Aware Opportunistic Routing (CAOR) algorithm. In [15], two social metrics, namely centrality and community, based on real human mobility traces, are exploited in the design of a social-based forward algorithm, coined BUBBLE.

The second major research thrust is content dissemination and sharing in opportunistic mobile social networks [16, 4, 17, 18, 19, 20]. In [16], the authors develop an analytical model to analyze epidemic information dissemination in opportunistic mobile social networks. In [4], the authors study information dissemination in integrated cellular and opportunistic networks in an attempt to bridge the gap between user social aspects and wireless networking. However, unlike our work, they focus on integrated cellular and opportunistic networks and employ Markov chain modeling tools. In [17], the authors focus on multicast and propose a Social-Similarity-based Multicast Algorithm (Multi-Sosim) using nodes dynamic social features and a compare-split scheme to improve multicast efficiency in impromptu mobile social networks. In [18], the authors propose a socially-aware network-based content dissemination scheme which outperforms centralized infrastructure-based content dissemination and leverages the users' social characteristics, e.g., common interest and social ties, among others. In [19], the focus is on cellular traffic offloading whereby it proposes a framework for Traffic Offloading assisted by Social network services (SNS) via opportunistic Sharing, coined TOSS, to offload SNS-based cellular traffic via user-to-user sharing. In [20], the paper studies content dissemination in opportunistic social networks and shows that non-social nodes with high contact rate (rarely found in "temporal communities") efficiently disseminate content. In contrast, the model proposed in our work is similarity-centric and information-theoretic.

Unlike the previous two thrusts, this work focuses on knowledge exchange in opportunistic mobile social networks. It is centered around user behavioral profiles and pair-wise similarity.

Mobile user profiles proposed in the literature can be grouped based on different perspectives. Few are based on user location, e.g., [21, 5], while others extend the profile to capture facets beyond location, e.g., [22, 23]. From another perspective, profiles may be classified into vector (nontemporal) and matrix (temporal) profiles depending on whether the temporal dimension is captured or not. Similarity assessment depends on the profile type and application context. Classic metrics exist for vector-based profiles such as cosine and Pearson correlation [24]. Distance metrics from probability theory, e.g., Hellinger distance [25], can be employed to test the similarity of probability distribution functions, like the user profiles proposed here. On the contrary, very few metrics are introduced for temporal profiles, e.g., singular value decomposition (SVD) based metrics [5].

In [26], the authors present a universal definition of similarity in terms of information theory. However, the similarity measure is derived from a set of assumptions rather than being directly stated as in earlier definitions. Moreover, experiments are not conducted on real user profiles to study the behavior of the proposed metrics. Examples of information-theoretic based models include cooperative data compression and distributed source coding for data collection in sensor networks with spatial correlations, e.g., [27, 28, 29, 30]. However, the main objective in that line of research is to eliminate redundancies among correlated sensor measurements [28, 29]. The joint entropy of the individual sensor random variables constitutes a lower bound on the total traffic volume generated

by the sensors that compression algorithms try to attain. On the contrary, our objective here is completely different. With the aid of information-theoretic constructs, we draw fundamental limits on the maximum knowledge available for a user in an opportunistic setting [31]. Later, we define the knowledge limit of a user as the upper bound on the knowledge a user can gain in an opportunistic encounter, characterized mathematically by the joint entropy of the knowledge users have. Moreover, nodes are generally fixed in sensor networks and the communications are largely multi-hop. On the other hand, in our problem setting, users are mostly mobile and communications are single-hop, yet, users can forward knowledge acquired over past single-hop encounters.

Recommendation services in the context of the Internet of Things (IoT) have been recently studied in [6]. However, the prime focus in [6] is to share data across IoT vertical applications for recommendation services, unlike this work which proposes the novel concept of opportunistic D2D recommendation systems.

4. Pair-wise Mobile User Similarity

One of the classic problems in Computer Science is similarity assessment, e.g., data mining, clustering and classification [32, 33]. For instance, it has received considerable attention for online social networks' recommendation systems in [34, 35, 36, 37]. In [34], the authors propose a model to infer the strength of relationships based on similarity and interactions. In [35], the similarity of users' ratings of items is computed using heuristic measures such as cosine similarity and Pearson correlation. The similarity is also studied in various contexts, e.g., web users recommendation [36] and peer recommendation systems [37].

In mobile scenarios, similarity has received limited attention through exploiting the users' spatio-temporal proximity (i.e., residing at the same place at the same time), e.g. [38, 39, 40]. In [40], similar users exchange ratings about touristic places they have previously visited. In [38] and [39], users can check who is in proximity and based on common interests may decide to establish communications. To the best of the authors' knowledge, mobile users' similarity has been investigated only in [38, 39, 41, 42]. However, the proposed mobile user profile is defined in terms of the geographic location only.

4.1. Proposed Mobile User Profiles

In this section, we propose a new model for profiling mobile users, that incorporates facets beyond the geographic location. In addition, we scrutinize non-temporal and temporal profiles. We assume V general life categories, e.g., books, travel, sports, among others, decided by the system designer taking into consideration the application of interest.

Thus, the non-temporal profile is a 1xV row vector where each element, C_i, represents the percentage of time spent by the mobile user, possibly online (*Interests*) or physical site visits (*Experiences*), in category i [43]. This vector is considered a probability mass function for the random

variable associated with the profile since $\sum_{i=1}^{V} C_i = 1$. The probability distribution definition of the user profile is not only convenient but also opens room for powerful mathematical tools to study similarity and knowledge sharing as discussed in Section 5. It is worth noting that, in general, the user profile design may take different forms and include different granularities, e.g., sports in general (as one category) to specific types of sports as sub-categories (e.g., football, basketball, swimming, etc.) depending on the user needs and application requirements. In this paper, we define the user profile, generically, as the probability distribution of user interests and experiences over a pre-specified set of life categories. The mathematical framework and tools in the paper are general enough to accommodate any set of life categories, which can be arbitrarily defined by the system/application designer. Finally, the detailed design of the profile, number and type of categories included as well as sub-categories is an important topic of research. However, it is left for future research, at this first look at the problem.

On the other hand, inspired by [5] which proposed a temporal profile matrix for the user Wi-Fi Access Point connectivity pattern and the key observation that simple vector profiles obscure paramount specifics about the user's temporal dynamics [43], we introduce probability distribution temporal profiles that capture facets other than location. Thus, profile vectors are obtained for K time windows where a window represents a day, week, etc. based on the user behavior dynamics and the target time horizon. This, in turn, yields a KxV matrix where the K profile vectors are the rows. Choosing the value of time granularity and the length of the horizon, K, is a pivotal issue which calls for applying data analytics tools on real-life traces capturing the users' behavior dynamics over time. This lies beyond the scope of this work. For our comparative analysis, we rely on real smartphone traces from the LiveLab project at Rice University [8] where the time window spans one day and $K = 197$ days on the average.

Given the proposed PMF user profiles, we move next to similarity assessment.

4.2. Similarity Metrics

The use of similarity metrics highly depends on the profile structure. For non-temporal profiles, cosine and Pearson correlation are widely used in the literature [24] with ranges $[0, 1]$ and $[-1, 1]$, respectively. These metrics are widely used due to their simplicity.

Motivated by the probability distribution definition of the proposed profile vectors, we explore distance metrics from probability theory, namely the Hellinger distance, Canberra distance and Jensen Shannon Divergence [25]. The Hellinger distance is defined for two probability mass functions (PMFs), A and B, as [25]

$$H(A, B) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{V} (\sqrt{a_i} - \sqrt{b_i})^2},$$

where $H(A, B) \in [0, 1]$ and Hellinger similarity is defined as $Sim_{HL}(A, B) = 1 - H(A, B)$.

On the other hand, the Canberra distance and Jensen-Shannon Divergence were problematic in our problem context since they yield infinite distance if one (or more) category in the profile vector is zero-valued. Zero-valued categories are typical in practice and were found to be recurrent in real-life traces, e.g., [8], where users' interests are clustered in few categories.

As for temporal profiles, we study two similarity metrics. First, a metric based on Singular Value Decomposition (SVD) from linear algebra proposed in [5]. Second, we propose a novel, low-complexity vectorized cosine metric that is motivated by the limitations of SVD. SVD is an extension to classic cosine similarity and is defined for two profile matrices X and Y as

$$Sim_{SVD}(X,Y) = \sum_{i=1}^{Rank(X)} \sum_{j=1}^{Rank(Y)} w_{xi} w_{y_j} |V_{X\,i}.V_{Y\,j}|, \quad (1)$$

which is, basically, the weighted cosine similarity between the two sets of eigen-behavior vectors, where $V_{X\,i}$ and $V_{Y\,j}$ are the ith and jth column of matrices V_X and V_Y, respectively. V_X and V_Y are the matrices resulting from the SVD transformation [44] of profile matrices X and Y, respectively, where $X = U_X \Sigma_X V_X^T$ and $Y = U_Y \Sigma_Y V_Y^T$.

On the positive side, SVD provides one provision for "anonymization' since the users exchange only the elements of Σ and V, but not matrix U. This, in turn, prevents eavesdroppers from reconstructing the sender's profile. On the downside, SVD similarity exhibits high computational complexity (proportional to the quadratic value of the history length K, for a fixed V). Furthermore, the similarity with oneself, $Sim_{SVD}(X,X)$, yields the maximum but not necessarily one, which causes problems while assessing similarity.

Motivated by the drawbacks of SVD and the simplicity of vector-based metrics, we propose a novel vectorized cosine (VCOS) metric with complexity scaling linearly with K. Thus, we convert the two KxV profile matrices, under investigation, to two 1xKV vectors using the vectorization operation in Linear Algebra [44] and then perform cosine similarity.

4.3. Similarity Metrics Comparison

In this section, we compare different similarity metrics using a real data set from the LiveLab Project at Rice University [8]. This data set consists of traces for smartphone users and Wi-Fi access points (APs) from 34 iPhone 3GS users. The 34 users included 24 Rice University students with traces spanning Feb. 2010 through Feb. 2011 and 10 Houston College students with traces from Sep. 2010 through Feb. 2011. The relevant data is stored in two tables. The first stores the names and genre (category) of 2500 iPhone Apps, as defined by the App Store. The Apps are classified into only 23 interest categories, e.g., business, travel, etc. The LiveLab data is carefully chosen since it hosts digital footprint logs for the mobile users in a categorized manner in contrast to other data sets in the literature which record only Wi-Fi AP connectivity traces that are irrelevant to our study. The second table covers the pattern by which each user accesses Apps on his/her phone. The pattern

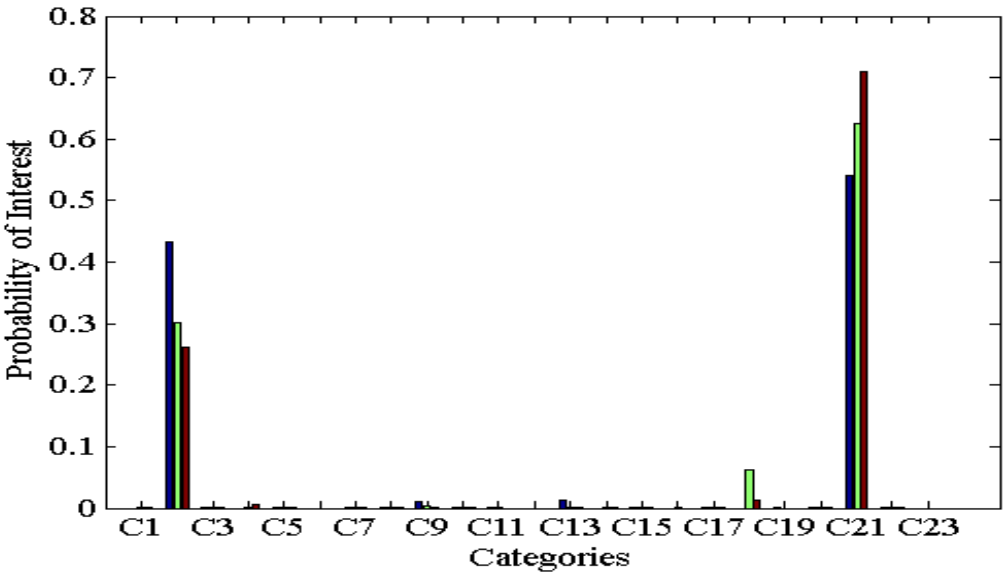


Figure 2: Three users' sample profiles from LiveLab smartphone traces.

consists of a history log for each of the 34 users with the date and access period. Cross-referencing the two tables, we build the non-temporal and temporal profiles for each user.

We observe that the resulting probability distribution profiles are zero-valued for most of the categories in most profiles and the user's interests lie in two to five categories as shown in Figure 2 and experienced in our real lives. This, in turn, makes the LiveLab users "qualitatively" similar. This key finding makes it impossible to use some metrics, such as Canberra Distance and Jensen-Shannon Divergence, with such sparse profiles due to the aforementioned infinite distance problem.

Thus, we focus on the cosine (COS), Hellinger (HLNG), SVD and Vectorized cosine (VCOS) similarity metrics[1] to evaluate the pairwise similarity of 34 LiveLab users, which yields 561 experiments. Figure 3 depicts the similarity outcomes of the four metrics vs the experiment index in a scatter plot. From Figure 3, we note the following:

- Cosine and Hellinger metrics yield higher similarity values in comparison with VCOS and SVD for the same pair of users. This confirms our intuition that temporal similarity metrics are "more thorough" and, therefore, conservative in declaring similarity. Thus, for a given threshold T between 0 and 1, two users may be declared "similar" using a non-temporal metric, yet, are deemed "dissimilar" using a temporal one. This is due to the fact that temporal profiles are generally more comprehensive than non-temporal metrics since they naturally do not filter out details and dynamics of users. This result is confirmed quantitatively in Table 1. The table

[1]Pearson correlation is not examined since it ranges from $[-1, 1]$ and mapping for comparison to other metrics skew the similarity results.

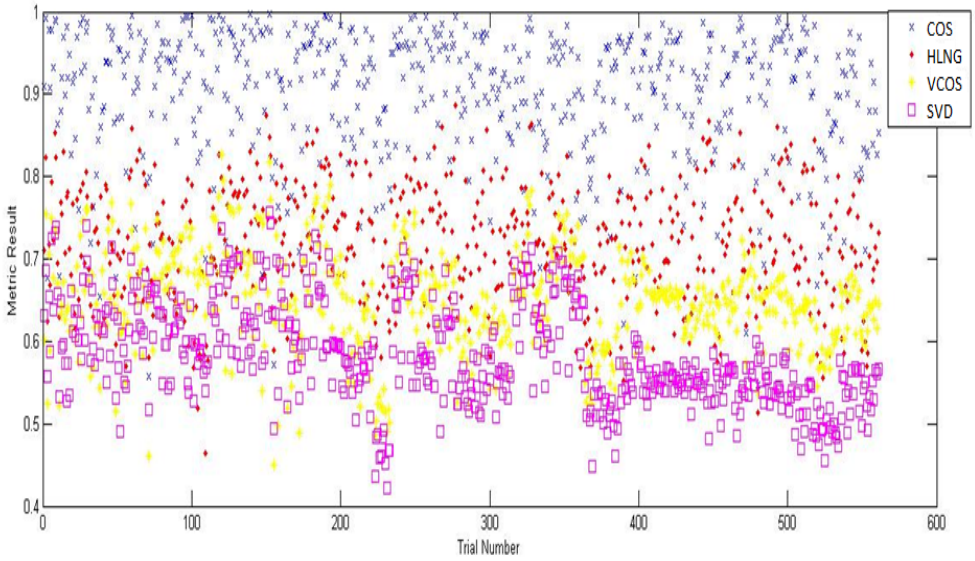


Figure 3: COS, HLNG, SVD and VCOS outcomes for pairwise similarity between LiveLab users.

Table 1: Comparison of the percentage of similar users under the four metrics for different similarity thresholds (T) for the LiveLab dataset.

$T =$	$[0.1, 0.4]$	0.5	0.6	0.7	0.8	0.9	1
SVD	100	92.51	34.41	3.57	0	0	0
VCOS	100	98.93	80.93	18.36	0.3565	0	0
HLNG	100	99.82	92.87	61.5	13.37	0	0
COS	100	100	99.47	97.33	89.13	59.18	0

confirms that VCOS and SVD yield a lower percentage of similar users than the non-temporal metrics (cosine and Hellinger). Hence, they are more cautious in declaring similar users.

- The Hellinger metric can be perceived to achieve a balance between the non-temporal and temporal paradigms since it is closest to the average of the four metrics [43]. Although this sheds some light and loosely shows the potential of Hellinger similarity, it still needs thorough analysis in future studies.

The metrics studied and proposed in this section, and the insights distilled constitute only a step towards answering the more challenging question of when are two users "actually similar", to serve as the ground truth in future work.

4.4. Scaling the results

To provide a larger interpretation of the prior analysis for the different similarity metrics, we complement our analysis with a large scale dataset (UF-WLAN) from a wireless local area network

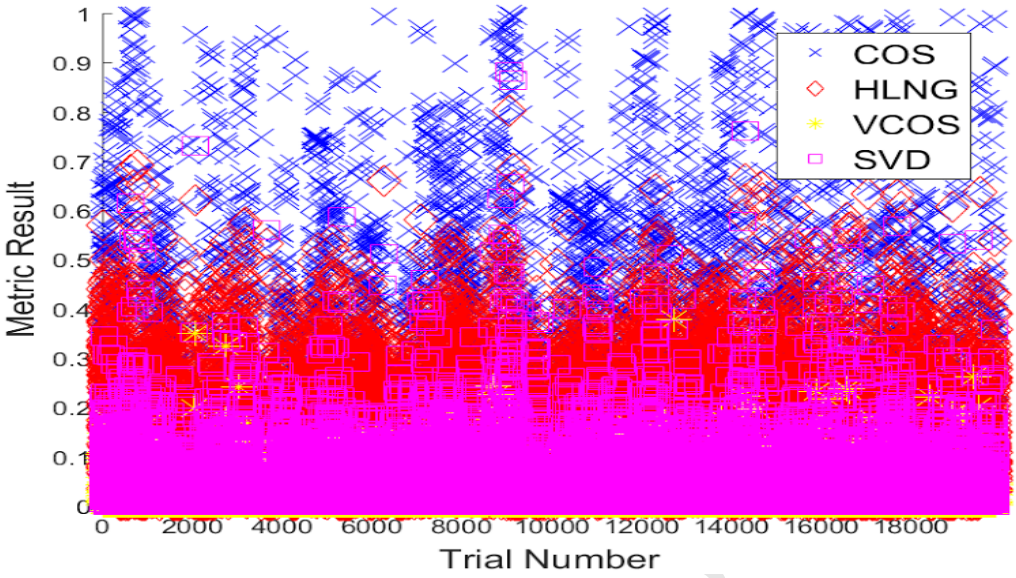


Figure 4: COS, HLNG, SVD and VCOS outcomes for pairwise similarity between UF-WLAN users.

(WLAN) at the campus of University of Florida. The WLAN logs contain wireless association and deauthentication events, for a period of 479 days in 2011-2012. Assuming MAC addresses are unique and unchanged, over 300,000 devices were online at least once. Each WLAN record provides a timestamp, assigned IP address at a corresponding access point (AP) and MAC address of the associated user device. Note that a user in the context of this study is defined as a single device. We cannot determine if two devices belong to the same person and could be correlated. Exact locations of the APs were not available, so their positions were approximated by the building locations where they were installed, i.e. the corresponding latitude/longitude returned by Google Maps API. To validate this approximation we fetched 8000 mapped APs around the campus area from the crowd-sourced service wigle.net. From 142 matched APs, in 58 buildings, all were within 200m or less from their mapped location. This error (1.5% of campus area) is reasonable considering the maximum AP coverage, inaccuracies in coarse-grained localization services and that we use the coordinates of the center of each building whereas users may see an AP on the edge of a building. The buildings are classified as academic, administrative, housing, library, museum, campus police, social, and sports. We avoid using these classes as the basis for user profiling since students will tend to cluster in academic, housing and library settings which would result in highly skewed and similar profiles. Instead we use the 142 buildings as categories for the previous user profile model. We produce non-temporal profiles for a set of randomly sampled 200 users from the most active users. We also generate temporal profiles where each row represents a day rendering the size of the profile to be 479 x 142. Our sample of 200 users gives 19,900 pairs for which the similarity results are depicted in Figure 4 and Table 2.

Table 2: Comparison of the percentage of similar users under the four metrics for different similarity thresholds (T) for the UF-WLAN dataset.

$T =$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
SVD	13.6	2.6	0.8	0.3	0.1	0	0	0	0	0
VCOS	0.7	0.1	0	0	0	0	0	0	0	0
HLNG	58	27.1	9.4	2.3	0.5	0.1	0	0	0	0
COS	37.4	22.2	14.4	9.7	6.2	3.7	2.2	1.2	0.6	0

The results presented in Figure 4 and Table 2 confirm our previous observation that temporal metrics (SVD and VCOS) are more conservative than non-temporal metrics (COS and HLNG).

5. Modeling and Fundamental Limits of Knowledge Sharing

In this section, we focus on knowledge sharing between similar users. Our main focus is to introduce a novel mathematical framework and build fundamental limits. Developing efficient knowledge sharing schemes is not the prime focus of this paper, yet, an important direction for future work. This framework lays the theoretical foundation for studying delay-tolerant knowledge sharing policies in opportunistic networks.

In particular, we characterize, using information theory concepts, the "knowledge" a user can extract in an opportunistic encounter, coined knowledge gain (KG), and the maximum knowledge a user can reap, coined knowledge gain limit (KL). Modeling abstractions have been heavily used in the literature to study the formation, evolution and dynamic behavior of social networks. For example, graph-theoretic tools and random graph models have been extensively employed in social networks studies to formally model patterns of networking, homophily, and clustering as well as basic concepts like centrality and connectivity, e.g., [45, 46]. In addition, they have been used to model social networks formation and growth, e.g., [47, 48]. However, to the best of our knowledge, employing information theory to formally model knowledge sharing and forwarding in opportunistic social networks has not been studied in the open literature.

5.1. Network Model and Assumptions

The notion of a "network" here, that is, nodes exchanging information, is established solely based on pairwise similarity, according to Section 4. Thus, if a group of users in an opportunistic encounter are all dissimilar, then there is no network since no knowledge sharing will follow. The scenario of interest is the one that involves a subset of similar users which triggers tips exchange. Accordingly, we focus on a group of nodes where all nodes are pairwise similar. We adopt a wireless ad hoc network model to represent an opportunistic setting of M pair-wise similar users. We assume that the network is formed because a user is pair-wise similar to all other users in the network. Each user has its own non-temporal profile vector, or multiple row vectors across the temporal dimension,

defined as a probability distribution over different categories as described in Section 4.1. We assume each user hosts a table storing *tips* and recommendations (i.e. knowledge) to share with similar users, e.g., bestsellers, events of common interest, etc. Finally, all nodes leverage wireless communication technologies with fixed power to communicate (i.e. circular range transmission), e.g., Wi-Fi or Bluetooth, and, therefore, the multiple access problem is assumed to be resolved.

In this section, we pose two key questions related to knowledge fundamental limits and sharing policies:

1. For user i, what is the fundamental limit on (i.e. maximum) knowledge that can be reaped by this user in an opportunistic network setting?

2. For user i, what is the knowledge gain that the user can actually attain from pair-wise similar users, for a given knowledge sharing scheme?

Towards addressing these questions, we introduce, next, terminology and definitions.

5.2. Definitions: Quantifying Knowledge

We commence by defining the Knowledge Gain and Knowledge Limit concepts as follows.

Definition V.1. The Knowledge Limit (KL_i) for user i is defined as the maximum knowledge that can be reaped by user i from pair-wise similar users in the network.

Definition V.2. The Knowledge Gain (KG_i) for user i is defined as the knowledge user i can collect from pair-wise similar users, using a given knowledge sharing scheme.

As given in the above definitions, the Knowledge Limit for user i (KL_i) is a fundamental limit and, hence, constitutes an upper bound on the knowledge that user i can acquire from similar users in the network, regardless of the knowledge sharing scheme. On the other hand, the Knowledge Gain (KG_i) is defined as the knowledge that user i can collect from pair-wise similar users, using a specific knowledge sharing scheme. In general, a given knowledge sharing scheme may/may not attain the KL available for the user, as shown later in the paper. Thus, by definition, $KG_i \leq KL_i$. This bound is general, irrespective of the specific network setup, topology or number of nodes. Motivated by the user profile structure defined as a probability distribution, we indicate that probability and information theory tools would be suitable for analyzing such systems.

The next step is to formally define the knowledge gain in a two-user encounter. The tips and recommendations (typically stored in a table) of an arbitrary user are assumed to follow the same distribution as that user's profile. Although this assumption might seem somewhat strong, it has practical relevance and its motivation is two-fold. First, this modeling assumption is reasonable since, in practice, typical mobile users would tend to have more tips (knowledge) in life categories they are more interested in (i.e. spend more time on, either consuming content or site/event visits). Second, it is a convenient mathematical abstraction for tips which are typically stored in the form of a database on the mobile users device. This assumption renders the proposed mathematical model tractable and the problem nicely lends itself to powerful tools for quantifying the knowledge gain

and associated fundamental limits, grounded in information theory.

5.2.1. The Knowledge Gain for a Two-User Encounter

Among basic information-theoretic concepts, the Entropy of a discrete-valued random variable X, denoted as $H(X)$, is a central concept that represents a measure of the information borne by this random variable [49]. Since the user recommendations/tips have an identical distribution to the user profile, we model tips as a discrete random variable, X. Accordingly, $H(X)$ formally characterizes the information (or knowledge) [2] the user possesses. This, in turn, allows us to define mathematically the new concepts of KL and KG.

First, we consider the simple example of a two-user opportunistic encounter, i.e. the users are within the communication range of each other. The two users are assumed to have tips/recommendations probability mass functions, denoted X and Y. Assume users X and Y meet, opportunistically, and are found similar[3], in the sense of Section 4. Accordingly, they start exchanging tips. Based on information theory, we characterize the following types of "knowledge" quantities:

1. Tips kept by user X but not by user Y, defined as $H(X|Y)$.
2. Tips kept by user Y but not by user X, defined as $H(Y|X)$.
3. Tips kept by both users, characterized as $I(X;Y)$, the mutual information between X and Y .

Note that the first type of tips is the knowledge gain for Y while the knowledge gain for user X from Y is defined as

$$KG(X) = H(Y|X) = H(X,Y) - H(X) \tag{2}$$

where $H(X,Y)$ is the joint entropy of X and Y modeling the tips probability distributions. The last type of tips (carried by the two users), denoted by the mutual information $I(X;Y)$, is considered the "communication overhead" since it is transmitted over the air without any contribution to increasing the knowledge of user X or Y. This is in complete agreement with the reasonable assumption that mobile users do not have any prior information about the others' individual tips when they opportunistically meet and, hence, this communication overhead is unavoidable.

In this toy example, the knowledge limit for any of the two users is equal to the attainable knowledge gain.

5.2.2. The Knowledge Limit

Given the aforementioned definitions of the knowledge gain and limit for a two-user encounter, the KL definition for user X_1, without loss of generality, in an opportunistic setting with $M-1$

[2] We use the terms Information and Knowledge interchangeably in the sequel.

[3] We abuse notation and use tips PMFs, X and Y, to refer to the users.

users, who are similar to X_1, is generalized as follows:

$$KL(X_1) = H(X_1, X_2, X_3, ..., X_M) - H(X_1) \tag{3}$$

which can be written as

$$KL(X_1) = H(X_2|X_1) + H(X_3|X_2, X_1) + ... + H(X_M|X_{M-1}, ..., X_1). \tag{4}$$

Thus, (3) shows that the maximum knowledge user X_1 can acquire, via exchanging tips with users in the network, is merely the aggregate knowledge all users bear, after removing redundancies, which is essentially the joint entropy, $H(X_1, X_2, X_3, ..., X_M)$, less the knowledge user X_1 already has, denoted by $H(X_1)$. It is worth noting that the KL characterization in (3) and (4) above is general, holds for any network topology and does not depend on knowledge sharing schemes.

5.3. Fundamental Limits and Policies

Utilizing the previous definitions section, we seek to characterize the KL for a user in different scenarios as well as its KG, under two sample knowledge sharing strategies: i) Send my own tips, coined "*Send Mine Only*", whereby a user shares his/her own tips with a pair-wise similar, encountered user and ii) Forward my own tips and others, coined "*Forward Mine Plus Others*", whereby a user shares his/her own tips and forward tips collected previously in other encounters.

These two policies are simple examples to explain the framework introduced in this paper, however, the model lends itself to studying and analyzing more sophisticated and advanced sharing strategies. For example, a user could selectively forward her own tips with a "portion" of others' tips based on a criteria of choice. This introduces a new class of strategies that deserves careful analysis, to quantify the strengths and trade-offs, which is a subject of future research.

The next step is to characterize the fundamental KL and the KG attainable by the two previously mentioned sharing strategies, while varying network configurations. We consider two opportunistic connectivity scenarios, namely single- and multi-hop topologies as well as two mobility scenarios, namely fixed topologies in case of quasi-stationary users and time-varying topologies caused by the user's portability within the same area.

5.3.1. Fixed Topology, Similarity-based Opportunistic Networks

A. Single-hop Network Topologies

In this scenario, the users could be fixed, quasi-stationary or portable, yet, any node remains in range with all other nodes, i.e. one-hop away, at all times. For this setting, KL is easily characterized, as in (3), and the KG will attain the limit. This is intuitive since a node can take turns to exchange tips with all other nodes in range. Thus, for any node, "all" available knowledge can be acquired. Next, we establish this result for the *Send Mine Only* (SMO) strategy.

Proposition 1. For single-hop network topologies using the SMO strategy, any node achieves its knowledge limit.

Proof. We assume, without loss of generality, that node X_1 encounters other nodes in an increasing order of their IDs. Under SMO, the cumulative knowledge gain for X_1, $KG(X_1)$, after receiving tips from all other nodes, $X_2, X_3, X_4, ..., X_M$ in turn, is given by $H(X_2|X_1) + H(X_3|X_2, X_1) + ... + H(X_M|X_{M-1}, ..., X_1)$, which is the same as $KL(X_1)$ in (4). The proposition is proven using a similar argument for all other nodes in the network. □

As indicated earlier, one of the key issues in this context is the amount of time it takes a user to achieve its KL, if attainable. This is directly related to the number of exchanges needed to attain the KL. Under SMO for single-hop networks, as in Proposition 1, and assuming that each node has one or more unique tips to contribute to the "network knowledge", then the worst-case (maximum) number of direct exchanges for an arbitrary user to attain the KL is simply $(M-1)$, i.e., $O(M)$.

Next, we quantify the KG of single-hop networks, under the Forward Mine Plus Others (FMPO) strategy. Thus, a user sends not only his/her own tips but also tips from previously encountered users, denoted by the subscript p. We prove in the next proposition that the KL is also attainable under FMPO.

Proposition 2. For single-hop network topologies using the *FMPO* strategy, any node achieves its knowledge limit.

Proof. We assume, without loss of generality, that each node initially has its own tips only and node X_1 encounters all other nodes in an increasing order of their IDs. The knowledge exchange goes over multiple rounds whereby in the first round, for instance, the following exchanges take place in parallel: $X_1 \leftrightarrow X_2$, $X_3 \leftrightarrow X_4$, $X_5 \leftrightarrow X_6$, etc. Thus, $KG(X_1)$ based on encountering nodes $X_2, X_3, X_4, ..., X_M$ in turn, is given by

$$\begin{aligned} KG(X_1) = {} & H(X_2, |X_1) + H(X_3, \vec{X_{3p}}|X_2, X_1) + H(X_4, \vec{X_{4p}}|X_3, \vec{X_{3p}}, X_2, X_1) + ... \\ & + H(X_M, \vec{X_{Mp}}|X_{M-1}, \vec{X_{(M-1)p}}, ..., X_1) \end{aligned} \tag{5}$$

where $\vec{X_{ip}}$ are the previous encounters of node X_i. After the first round, we notice that $X_{3p} = X_4$. The second round schedule would be $X_1 \leftrightarrow X_3$, $X_2 \leftrightarrow X_5$, and $X_4 \leftrightarrow X_6$. This would render $X_{4p} = X_6$, X_{6p}, X_3 where $X_{6p} = X_5$ yielding $X_{4p} = X_6, X_5, X_3$, and so on. Thus, substituting in (5) after $M/2$ rounds yields

$$\begin{aligned} KG(X_1) = {} & H(X_2, |X_1) + H(X_3, X_4|X_2, X_1) + H(X_4, X_3, X_5, X_6|X_4, X_3, X_2, X_1) + ... \\ & + H(X_M, \vec{X_{Mp}}|X_{M-1}, \vec{X_{(M-1)p}}, ..., X_1) \end{aligned} \tag{6}$$

Using the chain rule of entropies and expanding all terms in (6) yields some zero terms due to acquiring the same (redundant) knowledge from previous encounters. This, in turn, results in the KL formula in (4) and proves the proposition. □

It is worth noting that when the conditioning, in the conditional entropy terms in the RHS of (6), includes all nodes in the network, the incremental gain becomes zero and the node achieves its KL. Compared to SMO, FMPO attains the KL faster essentially due to the role of the previous encounters (appearing in the conditional entropy terms), which will be shown in Section 5.4. Evidently, this speed up is not for free due to an inherent trade-off between the cumulative KG build up over time on one hand and the incurred communication overhead on the other hand, which deserves further attention in future work, especially in multi-hop scenarios. The following proposition asserts that the communication overhead of FMPO is greater than or equal to SMO, in single-hop networks.

Proposition 3. For single-hop network scenarios, the communication overhead incurred under FMPO is greater than or equal to SMO.

Proof. We consider an encounter between two users, X and Y. Generalizing to a sequence of encounters is straightforward. The previous encounters for users X and Y are denoted by vectors $\vec{X_p}$ and $\vec{Y_p}$, respectively.

The communication overhead that user X incurs is the mutual information with user Y, that is, the knowledge overlap between what X sends (its own knowledge only in case of SMO) and Y's accumulated knowledge so far. Under SMO, the overhead is given by

$$OH(X)_{SMO} = I(X; Y, \vec{Y_p}). \tag{7}$$

Similarly, the overhead for Y is $OH(Y)_{SMO} = I(Y; X, \vec{X_p})$.

On the other hand, the communication overhead under FMPO is the same for both users as given below

$$OH(X)_{FMPO} = OH(Y)_{FMPO} = I(X, \vec{X_p}; Y, \vec{Y_p}). \tag{8}$$

It is well-known from information theory that the mutual information between random variables A and B is given by

$$I(A; B) = H(A) + H(B) - H(A, B). \tag{9}$$

Using (9) in (7) gives

$$OH(X)_{SMO} = H(X) + H(Y, \vec{Y_p}) - H(X, Y, \vec{Y_p}). \tag{10}$$

Using (9) in (8) gives

$$OH(X)_{FMPO} = OH(Y)_{FMPO} = H(X, \vec{X_p}) + H(Y, \vec{Y_p}) - H(X, \vec{X_p}, Y, \vec{Y_p}). \tag{11}$$

By subtracting (10) from (11), we get

$$OH(X)_{FMPO} - OH(X)_{SMO} = H(X, \vec{X_p}) - H(X) - H(X, \vec{X_p}, Y, \vec{Y_p}) + H(X, Y, \vec{Y_p}) \tag{12}$$

Based on the definition of joint entropy, $H(A, B) = H(A) + H(B|A) = H(B) + H(A|B)$, (12) can be written as

$$\begin{aligned} OH(X)_{FMPO} - OH(X)_{SMO} = H(X) + H(\vec{X_p}|X) - H(X) \\ - [H(\vec{X_p}|X, Y, \vec{Y_p}) + H(X, Y, \vec{Y_p})] + H(X, Y, \vec{Y_p}), \end{aligned} \tag{13}$$

which can be reduced to

$$OH(X)_{FMPO} - OH(X)_{SMO} = H(\vec{X_p}|X) - H(\vec{X_p}|X, Y, \vec{Y_p}) \geq 0, \tag{14}$$

where the inequality in (14) holds since conditioning does not increase entropy. This proves the sought result. □

B. Multi-hop Fixed Network Topologies

In this scenario, the network topology is time-invariant and always connected where some nodes are multi-hop away from each other. In this setting, the effect of knowledge sharing strategies prevails and plays a key role in whether a user can or cannot, achieve the knowledge limit.

Next, we investigate the KL achievability and trade-offs for SMO and FMPO under fixed multi-hop networks. The knowledge gain achieved by node X_1, in case of SMO, is limited by the size of the single-hop neighborhood ($N < M$) which results in a KG strictly less than KL. We formally prove this result in the next proposition.

Proposition 4. For multi-hop fixed network topologies, SMO is not guaranteed to achieve the knowledge limit, i.e., $KG(X_1) \leq KL(X_1)$ iff $N < M$.

Proof. We assume, without loss of generality, that node X_1 communicates with other nodes in an increasing order of their IDs. The cumulative KG for node X_1, $KG(X_1)$, according to exchanges with neighbors $X_2, X_3, X_4, ..., X_N$ is given by $H(X_2|X_1) + H(X_3|X_2, X_1) + ... + H(X_N|X_{N-1}, ..., X_1)$. Notice that the sum of non-negative conditional entropy terms is limited to the single-hop neighborhood, $N < M$ nodes, which misses other non-negative terms involving the $M - N$ non-neighbors of X_1. Hence, it follows that $KG(X_1) \leq KL(X_1)$, which establishes the proof. □

It is worth noting that the special case of $N = M$, for all nodes, reduces to the single-hop network case where we have shown in Section 5.3.1.A that the KL is achievable under both knowledge sharing policies. An interesting, and somewhat surprising insight, which will be discussed later, is that nodes mobility can be leveraged to achieve the knowledge limit, even if $N < M$.

We shift our attention next to the performance of FMPO for multi-hop fixed network topologies. As predicted, forwarding tips acquired from other users could allow a node to achieve its knowledge limit, even if $N < M$. We prove this result in the following proposition.

Proposition 5. For multi-hop fixed network topologies, any node can achieve the knowledge limit using the FMPO strategy.

Proof. We assume, without loss of generality, that each node is initialized with its own knowledge only and user X_1 goes through exchanges with single-hop neighbors in an increasing order of their IDs. At the same time, other neighbors proceed with pairwise encounters with other users in the network. The cumulative KG for X_1, $KG(X_1)$, after meeting single-hop neighbors $X_2, X_3, X_4, ..., X_N$ is given by

$$\begin{aligned} KG(X_1) = H(X_2, |X_1) + H(X_3, \vec{X_{3p}}|X_2, X_1) + H(X_4, \vec{X_{4p}}|X_3, \vec{X_{3p}}, X_2, X_1) + ... \\ + H(X_N, \vec{X_{Np}}|X_{N-1}, \vec{X_{(N-1)p}}, ..., X_1) \end{aligned} \tag{15}$$

Next, we have two cases. First, if the previous knowledge vectors, namely $\vec{X_{ip}}$ $\forall i$, carry all knowledge (forwarded tips) from non-neighboring nodes, namely $X_{N+1}, X_{N+2}, ..., X_M$, then it is straightforward to show that the cumulative KG of X_1 is

$$KG(X_1) = H(X_1, X_2, X_3, ..., X_M) - H(X_1) = KL(X_1) \tag{16}$$

which proves the result. Second, if previous encounters for neighbors of X_1 do not cover knowledge from all non-neighbors in the network, then node X_1 would still need to re-encounter its single-hop neighbors to achieve $KL(X_1)$. Backed by network connectedness and unconstrained delay, user X_1 will achieve its KL almost surely via repeated pairing with single-hop neighbors it has paired with before (to reap new knowledge they acquired over time) until it acquires all missing knowledge from non-neighboring nodes beyond its radio range. This proves the result. □

5.4. User Traces and Numerical Results

In this section, we supplement our analytical findings so far with performance results based on real smartphone profile traces [8] and real-life user mobility traces, gathered at Infocom 2005 [9, 10].

5.4.1. Single-hop Network Topologies

Our experiments incorporate real traces, either for user profiles or mobility. For user behavior, we utilize digital footprint traces (interests) for 20 smartphone users, over $V = 24$ life categories, from the LiveLab project [8]. In order to quantify the KL and KG for a user, we need to process a huge six month worth of interests data. To this end, we determine the joint probability for the 20 profiles under investigation in two steps as follows.

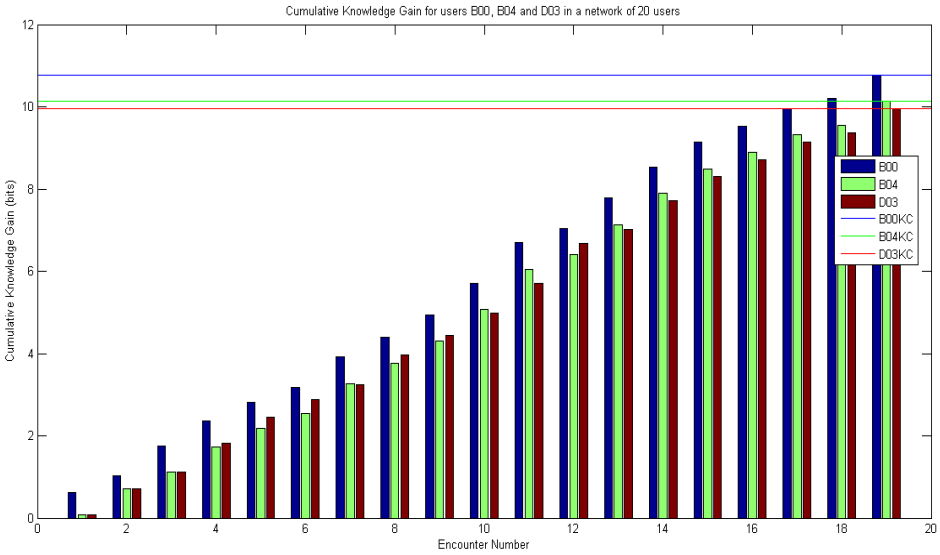


Figure 5: Cumulative KG over time for three arbitrary users in a single-hop network topology under SMO.

- Step 1: record the users' activities categorized under 24 categories[4], each second, and record their concurrent activities over the period from September 2010 to February 2011.
- Step 2: normalize the results of the first step over the whole six month duration to get the joint PMF.

We show next that more node encounters over time results in increasing the cumulative knowledge gain.

First, we give results for a single-hop network under SMO. For the $M = 20$ nodes network described earlier, any user can achieve the KL within $M - 1 = 19$ encounters. This is confirmed in Figure 5 for three arbitrary users, namely $B00$, $B04$ and $D03$. As depicted in Figure 5, the cumulative KG is a non-decreasing function over time. The KL, on the other hand, is a horizontal line that generally varies from one user to another, depending on the user's own knowledge and the amount of prior knowledge the user bears before encounters others.

Next, we consider the same network topology, yet, under FMPO knowledge sharing. Grounded in Proposition 2, we confirm that for single-hop networks, all nodes attain the KL using the FMPO strategy, yet, in fewer encounters compared to SMO, due to forwarding the tips of others. We confirm this behavior for three arbitrary users, namely $B00$, $B04$ and $D03$, in Figure 6.

[4]The 24th category captures the case when the smartphone is off or not running any application.

5.4.2. Multi-hop Fixed Network Topologies

Recall from Proposition 4 that attaining the KL of any user using SMO is limited by the single-hop neighborhood size of this node, denoted N. Therefore, we examine 100 randomly generated topologies of users, uniformly distributed, whereby a user has only $6-7$ single-hop neighbors, on the average. It should be noted that user $B00$ does not achieve the knowledge limit, as established in Proposition 4 and confirmed using real smartphone behavior traces in Figure 7. Thus, the maximum knowledge gain attainable by node $B00$ is only 46.33% of its knowledge limit. Similarly, users $B06$ and $D00$ have single-hop neighborhoods strictly less than the network size and, hence, cannot achieve their respective knowledge limits using SMO.

To conclude our fixed networks performance evaluation, we analyze the KG and KL performance of the FMPO policy in multi-hop topologies. In this case, the FMPO policy is expected to overcome the limited neighborhood problem due to sharing others' tips and, hence, nodes could attain the knowledge limit as proven before in Proposition 5. The results here are based on 100 randomly generated topologies. The cumulative knowledge gains for users $B00$, $B06$ and $D00$ are depicted in Figure 8. We notice that the KL is achievable for the three shown users after 7 encounters for B00 and B06, and 8 encounters for D00.

5.4.3. Time-varying Topology (Mobile) Networks

A. User Profiles and Mobility Traces

After an extensive search for mobile user traces on publicly available data repositories, e.g., CRAWDAD [9, 10] and the alike, we did not find traces that include both, user behavior and mobility

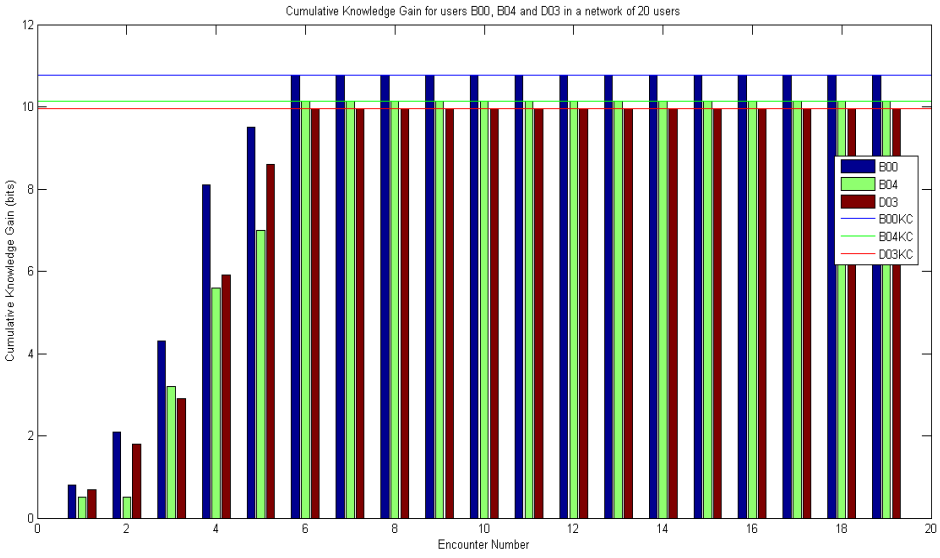


Figure 6: Cumulative KG over time for three arbitrary users in a single-hop network topology under FMPO.

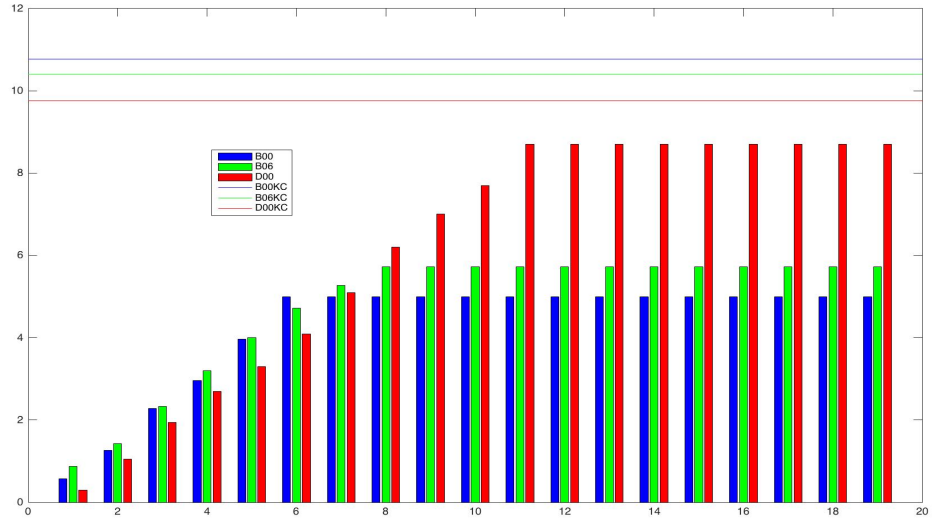


Figure 7: Cumulative KG over time for three arbitrary users in a multi-hop fixed network topology under SMO.

Figure 8: Cumulative KG over time for three arbitrary users in a multi-hop fixed network topology under FMPO.

traces. Moreover, most of the mobility traces are university campus Wi-Fi access patterns as opposed to mobile user encounters. In order to proceed with the performance evaluation based on real data, we resort to jointly leveraging traces from two different data sets, for user behavior and mobility. First, user profiles are constructed based on the LiveLab project data [8] described earlier. On the other hand, mobility traces are based on a "conference encounter" data, namely Infocom 2005 [9, 10]. For the Infocom 2005 experiment, the data set is relatively small whereby participants are 50 attending the student workshop. Nevertheless, it constitutes a reasonably sized set for our performance evaluation purposes. The students were given iMotes on March 7^{th}, 2005 between lunch time and 5 pm and collected on March 10th, 2005 in the afternoon. Two iMotes were lost while seven did not deliver useful data due to an accidental hardware reset. Contacts with these nine iMotes were discarded from the traces of others to avoid any effect on the results. The first six hours are discarded since they were attending the same workshop. We consider the contacts of 20 nodes only to match the number used from the LiveLab user profiles data. Thus, we associate the profiles of 20 randomly chosen users from the LiveLab data set to the mobility traces of 20 iMotes from Infocom 2005 and monitor them for half a day. This enables us to conduct our knowledge sharing analysis and collect the sought performance results.

Despite the fact that user profile and mobility traces are brought from two totally independent data sets, we find it a very useful attempt towards evaluating our policies, due to the lack of the sought data in the public domain. This constitutes a strong motivation for the mobile networking and computing community to focus on the social dimension as well as the mobility and wireless connectivity dimensions, which already have several data sets in the public domain.

B. Performance Results

In this section, we quantify the knowledge limit and gain of time-varying topology (mobile) multi-hop networks, under the two sharing policies. Intuitively, users' mobility would play a key role in whether a node can achieve its KL and, if so, how much time this would incur. The gathered results are shown in Table 3. We compare the KG acquired by sample nodes using SMO in two cases, namely the stationary case where a snapshot is taken at time $t = 0$ and the mobile case over half a day. At time, $t = 0$, all nodes, except for node $B07$, are disconnected yielding KG of zero. Node $B07$ is initially connected to $D06$ and reaps a KG of 0.64 as shown. The intriguing observation here is that mobility does help some nodes approach their knowledge limit, e.g., $B00, B03, B05, B07, B08, B09$. On the other hand, some nodes, e.g., $B06$, do not benefit from mobility since they remain disconnected throughout the experiment lifetime. This insight agrees with intuition since the mobility patterns of some nodes could assist them in encountering the "knowledge hotspots" of the network and acquiring knowledge faster than others. On the other hand, the mobility of other nodes could give rise to encounters with very slim/no KG benefits, e.g. nodes $B02$ and $B06$. Finally, we highlight that FMPO achieves KG no less than SMO, over the same period of time, which agrees with our

theoretical findings.

Extensive studies of user encounter patterns in campus WLANs, e.g., [50, 10], have shown that, on the average, a user encounters only 2% of the population in a month and pointed out the heavy clustering of a user's behavior (spending 90% of their online time within only five APs (out of 600). In the following proposition, we establish the result based on an ideal mobility model guaranteeing encounters with all other nodes, which may not be valid for a whole campus scenario according to to [50]. Nevertheless, for local encounters and mobile communities, the encounter ratio tends to be quite high (vs. 2% for the whole population) and our model is likely to be valid for realistic mobility scenarios.

Based on the seminal work on the effect of mobility on the throughput and delay in wireless ad hoc networks [51], the following proposition proves that the knowledge limit in mobile, delay-tolerant, multi-hop social networks is always achievable, under idealistic assumptions and loose delay constraints. Under those assumptions, an arbitrary node will encounter all other nodes in the network, almost surely. Nevertheless, modeling realistic mobility and characterizing the conditions under which the KG is improved by mobility is an interesting subject of future research.

Proposition 6. For a time-varying topology network, an arbitrary node achieves its KL under loose delay constraints, almost surely, in case each node moves according to an independent two-dimensional random walk in a fixed area.

Proof. In case of loose delay constraints and independent two-dimensional random walks in a fixed area, it has been shown in the literature that an arbitrary node will encounter all other nodes in the network, almost surely (refer to Lemma 6 in [51]). Hence, we assume without loss of generality that node X_1 has exchanges with all nodes in the network, when it encounters them, in an ascending order of their node IDs. Using SMO, the cumulative knowledge gain for node X_1, $KG(X_1)$, based on encountering nodes $X_2, X_3, X_4, ..., X_M$ is the same as (4). Similar arguments can be employed to prove the same result using the FMPO policy, which proves the proposition. □

Table 3: Cumulative knowledge gain (in bits) for nine mobile users after half a day.

Users	KG using SMO for stationary nodes (t=0)	KG using SMO	KG using FMPO	KL
B00	0	7.12	9.12	10.76
B02	0	0.63	9.05	10.24
B03	0	7.44	7.93	10.44
B04	0	6.94	7.62	10.13
B05	0	9.03	9.03	10.22
B06	0	0	0	10.4
B07	0.64	8.82	8.82	10.46
B08	0	7.78	8.45	10.09
B09	0	8.36	8.97	10.34

6. Conclusion

We propose in this paper a novel mathematical framework for similarity-based opportunistic social networks. We first propose generalized, non-temporal and temporal profiles as a probability mass function. Second, we study classic and information-theoretic metrics for similarity using public domain data. We conclude that temporal metrics result in lower similarity indices compared to non-temporal metrics, on the average, due to capturing details and behavioral variations in the temporal dimension. Third, we introduce a novel information-theoretic framework for knowledge sharing among similar, opportunistic users. Finally, we present performance results quantifying the cumulative knowledge gain over time and its upper bound, the knowledge limit, using public domain traces for user behavior and mobility, in case of fixed and mobile scenarios.

The promising research direction explored in this paper is still ripe and opens ample room for future research and can be extended along a number thrusts. For instance, establishing trust between opportunistic users is a key enabler for opportunistic D2D services. Second, proposing novel similarity metrics which capitalize on the strengths and insights of non-temporal and temporal profiles pointed out in this paper. Third, an extensive analysis for the Hellinger and vectorized cosine similarity metrics, with diverse user communities and datasets is expected to deepen the community understanding of these new similarity metrics and could give rise to other novel metrics as well. Fourth, leverage the proposed framework to analyze novel and efficient knowledge sharing policies. Finally, establish fundamental limits and study the effect of diverse user mobility patterns on knowledge sharing and emerging opportunistic D2D services.

References

[1] I. T. Union, ITU releases 2016 ICT figures, http://www.itu.int/en/mediacentre/Pages/2016-PR30.aspx, accessed: 04/28/2017 (July 2016).

[2] I. Smith, Social-mobile applications, Computer 38 (4) (2005) 84–85.

[3] Y.-J. Chang, H.-H. Liu, L.-D. Chou, Y.-W. Chen, H.-Y. Shin, A general architecture of mobile social network services, in: International Conference on Convergence Information Technology, IEEE, 2007, pp. 151–156.

[4] J. Hu, L. Yang, V. Poor, L. Hanzo, Bridging the social and wireless networking divide: Information dissemination in integrated cellular and opportunistic networks, IEEE Access 3 (2015) 1809–1848.

[5] W. Hsu, D. Dutta, A. Helmy, CSI: A paradigm for behavior-oriented profile-cast services in mobile networks, Ad Hoc Networks 10 (8) (2012) 1586–1602.

[6] Y. Saleem, N. Crespi, M. Rehmani, R. Copeland, D. Hussein, E. Bertin, Exploitation of social iot for recommendation services, in: IEEE 3rd World Forum on Internet of Things (WF-IoT), IEEE, 2016, pp. 359–364.

[7] S. Trifunovic, F. Legendre, C. Anastasiades, Social trust in opportunistic networks, in: IEEE Conference on Computer Communications (INFOCOM), IEEE, 2010, pp. 1–6.

[8] C. Shepard, A. Rahmati, C. Tossell, L. Zhong, P. Kortum, Livelab: measuring wireless networks and smartphone users in the field, ACM SIGMETRICS Performance Evaluation Review 38 (3) (2011) 15–20.

[9] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, A. Chaintreau, CRAWDAD trace set cambridge/haggle/imote (v. 2009-05-29), Downloaded from http://crawdad.org/cambridge/haggle/20090529/imote (may 2009).

[10] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, J. Scott, Impact of human mobility on opportunistic forwarding algorithms, IEEE Transactions on Mobile Computing 6 (6) (2007) 606–620.

[11] N. Eagle, A. Pentland, D. Lazer, Inferring friendship network structure by using mobile phone data, Proceedings of the National Academy of Sciences of the United States of America 106 (36) (2009) 15274–15278.

[12] M. ElSherief, T. ElBatt, A. Zahran, A. Helmy, O'BTW: an opportunistic, similarity-based mobile recommendation system, in: Proceeding of the 11th ACM annual international conference on Mobile systems, applications, and services (MobiSys), ACM, 2013, pp. 511–512.

[13] N. Vastardis, K. Yang, Mobile social networks: Architectures, social properties, and key research challenges, IEEE Communications Surveys & Tutorials 15 (3) (2013) 1355–1371.

[14] M. Xiao, J. Wu, L. Huang, Community-aware opportunistic routing in mobile social networks, IEEE Transactions on Computers 63 (7) (2014) 1682–1695.

[15] P. Hui, J. Crowcroft, E. Yoneki, Bubble rap: Social-based forwarding in delay-tolerant networks, IEEE Transactions on Mobile Computing 10 (11) (2011) 1576–1589.

[16] Q. Xu, Z. Su, K. Zhang, P. Ren, X. S. Shen, Epidemic information dissemination in mobile social networks with opportunistic links, IEEE Transactions on Emerging Topics in Computing 3 (3) (2015) 399–409.

[17] Y. Xu, X. Checn, Social-similarity-based multicast algorithm in impromptu mobile social networks, in: IEEE Globecom, IEEE, 2014, pp. 346–351.

[18] J. Hu, L. Yang, K. Yang, L. Hanzo, Socially aware integrated centralized infrastructure and opportunistic networking: a powerful content dissemination catalyst, IEEE Communications Magazine 54 (8) (2016) 84–91.

[19] X. Wang, M. Chen, Z. Han, D. Wu, T. Kwon, Toss: Traffic offloading by social network service-based opportunistic sharing in mobile social networks, in: IEEE INFOCOM, IEEE, 2014, pp. 2346–2354.

[20] A.-K. Pietiläinen, C. Diot, Dissemination in opportunistic social networks: the role of temporal communities, in: ACM Mobihoc, ACM, 2012, pp. 165–174.

[21] W. Hsu, D. Dutta, A. Helmy, Profile-cast: Behavior-aware mobile networking, in: Wireless Communications and Networking Conference (WCNC), IEEE, 2008, pp. 3033–3038.

[22] J. Ramer, A. Soroca, D. Doughty, Mobile user profile creation based on user browse behaviors, US Patent App. 11/929,129 (Oct. 2007).

[23] S. Moghaddam, A. Helmy, S. Ranka, M. Somaiya, Data-driven co-clustering model of internet usage in large mobile societies, in: Proceedings of the 13th ACM international conference on Modeling, analysis, and simulation of wireless and mobile systems, ACM, 2010, pp. 248–256.

[24] W. Jones, G. Furnas, Pictures of relevance: A geometric analysis of similarity measures, Journal of the American Society for Information Science 38 (6) (1987) 420–442.

[25] S.-H. Cha, Comprehensive survey on distance/similarity measures between probability density functions, City 1 (2) (2007) 1.

[26] D. Lin, An information-theoretic definition of similarity, in: ICML, Vol. 98, Citeseer, 1998, pp. 296–304.

[27] S. Pradhan, K. Ramachandran, Distributed source coding: symmetric rates and applications to sensor networks, in: Data Compression Conference, IEEE, 2000, pp. 363–372.

[28] D. Marco, E. Duarte-Melo, M. Liu, D. Neuhoff, On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data, in: Information Processing in Sensor Networks, Second International Workshop, IPSN, Springer, 2003, pp. 1–16.

[29] S. Pattem, B. Krishnamachari, R. Govindan, The impact of spatial correlation on routing with compression in wireless sensor networks, ACM Transactions on Sensor Networks (TOSN) 4 (4) (2008) 24.

[30] T. ElBatt, On the trade-offs of cooperative data compression in wireless sensor networks with spatial correlations, IEEE Transactions on Wireless Communications 8 (5) (2009) 2546–2557.

[31] M. ElSherief, T. ElBatt, A. Zahran, A. Helmy, An information-theoretic model for knowledge sharing in opportunistic social networks, in: The 8th IEEE International Conference on Social Computing and Networking (SocialCom), IEEE, 2015, pp. 446–451.

[32] R. T. Ng, J. Han, Efficient and effective clustering methods for spatial data mining, in: Proc. of the 20th International Conference on Very Large Data Bases, 1994, pp. 144–155.

[33] P. Berkhin, A survey of clustering data mining techniques, in: Grouping multidimensional data, Springer, 2006, pp. 25–71.

[34] R. Xiang, J. Neville, M. Rogati, Modeling relationship strength in online social networks, in: Proceedings of the 19th ACM International Conference on World wide web, ACM, 2010, pp. 981–990.

[35] I. Konstas, V. Stathopoulos, J. M. Jose, On social networks and collaborative recommendation, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, 2009, pp. 195–202.

[36] M. Pazzani, A framework for collaborative, content-based and demographic filtering, Artificial Intelligence Review 13 (5-6) (1999) 393–408.

[37] A. Tveit, Peer-to-peer based recommendations for mobile commerce, in: Proceedings of the 1st ACM International Workshop on Mobile commerce, ACM, 2001, pp. 26–29.

[38] M. SIEGLER, Skout brings location-based dating to the iphone, `http://venturebeat.com/2009/01/21/skout-brings-location-based-dating-to-the-iphone/`, accessed: 01/04/2012 (January 2009).

[39] I. Trestian, S. Ranjan, A. Kuzmanovic, A. Nucci, Measuring serendipity: connecting people, locations and interests in a mobile 3g network, in: Proceedings of the 9th ACM SIGCOMM Internet measurement conference, ACM, 2009, pp. 267–279.

[40] A. De Spindler, M. Norrie, M. Grossniklaus, B. Signer, Spatio-temporal proximity as a basis for collaborative filtering in mobile environments, in: CAISE, Citeseer, 2006, pp. 912–926.

[41] M.-J. Lee, C.-W. Chung, A user similarity calculation based on the location for social network services, in: Database Systems for Advanced Applications, Springer, 2011, pp. 38–52.

[42] G. S. Thakur, A. Helmy, W.-J. Hsu, Similarity analysis and modeling in mobile societies: the missing link, in: Proceedings of the 5th ACM workshop on Challenged networks, ACM, 2010, pp. 13–20.

[43] M. ElSherief, T. ElBatt, A. Zahran, A. Helmy, The quest for user similarity in mobile societies, in: The 2nd International Workshop on Social and Community Intelligence (IEEE Percom Workshops), IEEE, 2014, pp. 569–574.

[44] G. Strang, The fundamental theorem of linear algebra, American Mathematical Monthly 100 (9) (1993) 848–855.

[45] L. C. Freeman, Centrality in social networks conceptual clarification, Social networks 1 (3) (1978) 215–239.

[46] J. A. Barnes, Graph theory and social networks: A technical comment on connectedness and connectivity, Sociology 3 (2) (1969) 215–232.

[47] P. Erdos, A. Renyi, On the evolution of random graphs, Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5 (1) (1960) 17–60.

[48] M. E. Newman, D. J. Watts, S. H. Strogatz, Random graph models of social networks, Proceedings of the National Academy of Sciences of the United States of America 99 (Suppl 1) (2002) 2566–2572.

[49] J. Thomas, T. Cover, Elements of Information Theory, 2nd Edition, John Wiley & Sons, Inc., 2006.

[50] W.-j. Hsu, A. Helmy, On nodal encounter patterns in wireless lan traces, IEEE Transactions on Mobile Computing 9 (11) (2010) 1563–1577.

[51] A. El Gamal, J. Mammen, B. Prabhakar, D. Shah, Throughput-delay trade-off in wireless networks, in: IEEE INFOCOM'04, IEEE, 2004, p. 475.

Mai ElSherief

Mai ElSherief is a Ph.D. candidate in the Computer Science department at the University of California, Santa Barbara (UCSB). She received her BSc. in 2011 in Computer Engineering from Cairo University, Egypt and a MSc. in 2013 in Wireless Communication from the School of Communication and Information Technology, Nile University, Egypt. In 2017, she received the Fiona and Michael Goodchild graduate mentoring award. In 2009, she received the Cairo University Ideal student award. Her research interests lie in data analytics, social computing, social networks, information science, and social good.

Babk Alipour

Babak Alipour is a Ph.D. student in the Computer and Information Sciences and Engineering department at University of Florida (UF). He received a B.Eng in 2014 from Information Technology from Amirkabir University of Technology (Tehran Polytechic), Iran. His research interests lie in big data analytics in mobile networks for modeling, simulation and benchmarking of protocols.

Mimonah AlQathrady

Mimonah Al Qathrady is a Ph.D. student in the Computer and Information Sciences and Engineering at the University of Florida (UF). She received a M.S degree in Computer Engineering 2013 from UF. She received the B.S in Information Systems 2007 from King Khalid University. Her current research interest includes indoor mobility modeling, infection tracing, IoT encounter distance and proximity estimation and classification.

Tamer ElBatt

Tamer ElBatt received the B.S. and M.S. degrees in EECE from Cairo University, Egypt in 1993 and 1996, respectively, and the Ph.D. degree in Electrical and Computer Engineering from the University of Maryland, College Park, MD, USA in 2000. From 2000 to 2009 he was with major U.S. industry R&D labs, e.g., HRL Laboratories, LLC, Malibu, CA, USA and Lockheed Martin ATC, Palo Alto, CA, USA, at various positions. From 2009 to 2017, he served at the EECE

Dept., Faculty of Engineering, Cairo University, as an Assistant Professor and later as an Associate Professor, currently on leave. He also held a joint appointment with Nile University, Egypt from 2009 to 2017 and has served as the Director of the Wireless Intelligent Networks Center (WINC) from 2012 to 2017. In July 2017, he joined the Dept. of CSE at the American University in Cairo as an Associate Professor. Dr. ElBatt research has been supported by the U.S. DARPA, ITIDA, QNRF, EU FP7, H2020, General Motors, Microsoft, Google and Vodafone Egypt Foundation and is currently being supported by Egypt NTRA. He has published more than 100 papers in prestigious journals and international conferences. Dr. ElBatt holds seven issued U.S. patents. Dr. ElBatt is a Senior Member of the IEEE and has served on the TPC of numerous IEEE and ACM conferences. He served as the Demos Co-Chair of ACM Mobicom 2013 and the Publications Co-Chair of IEEE Globecom 2012 and EAI Mobiquitous 2014. Dr. ElBatt currently serves on the Editorial Board of IEEE Transactions on Cognitive Communications and Networking and Wiley International Journal of Satellite Communications and Networking and has previously served on the Editorial Board of IEEE Transactions on Mobile Computing. Dr. ElBatt has also served on the United States NSF and Fulbright review panels. Dr. ElBatt was a Visiting Professor at the Dept. of Electronics, Politecnico di Torino, Italy in Aug. 2010, FENS, Sabanci University, Turkey in Aug. 2013 and the Dept. of Information Engineering, University of Padova, Italy in Aug. 2015. Dr. ElBatt is the recipient of the 2014 Egypt's State Incentive Award in Engineering Sciences, the 2012 Cairo University Incentive Award in Engineering Sciences and the prestigious Google Faculty Research Award in 2011. His research interests lie in the broad areas of performance analysis, design and optimization of wireless and mobile networks.

Ahmed Zahran

Ahmed H. Zahran is currently a Lecturer at the Department of Computer Science, University College Cork. He previously worked in different institutions including University College Cork (UCC), Ireland, Cairo University, Nile University (Egypt), University of Toronto. He obtained his Ph.D. at the Department of Electrical and Computer Engineering, the University of Toronto in 2007. He also received his BSc and MSc in Electrical Engineering from Electronics and Electrical Communication Department at Faculty of Engineering, Cairo University in 2000 and 2003 respectively. His research interests span different topics in wireless mobile networking, such as software-defined networks, multimedia streaming, cognitive networking, energy efficient networking, mobility management, traffic management, and resource management, and modeling and performance evaluation. His research received several awards including the first DASH Industry Forum Excellence in DASH Award at the ACM MMSys 2017 conference, best demo in IEEE LANMAN 2016, the UCC commercialization award 2010, the spotlight paper for IEEE Trans. Mobile Computing (June 2010), and the best paper award in IFIP Networking 2005 conference. He served as a technical reviewer for various journals and conferences including the IEEE Trans. on Mobile Computing, IEEE Transactions on Wireless Communications, IEEE Communication Mag. He also served on the technical program committee of several conferences and workshops.

Ahmed Helmy

Ahmed Helmy received his PhD in Computer Science in 1999 from the University of Southern California (USC), his MS in Electrical Engineering (EE) in 1995 from USC, his MS in Engineering Mathematics in 1994 and his BS in EE in 1992 from Cairo University. He was a key researcher in the NS-2 and PIM projects at USC/ISI from 1995–1999. Starting in 1999, he has been on the EE Department faculty at USC, and the director of the wireless networking lab. Since fall 2006, he has been a professor and the director of the mobile networking laboratory in the CISE Department at the University of Florida (UF). In 2017 he was the August-Wilhelm Scheer honorary fellow at the Technical University of Munich, in 2016 he was a visiting professor at UPMC, Paris, and in 2014 he was a visiting professor at Bogazici University, Istanbul. He received the NSF CAREER Award in 2002, the Zumberge Research Award in 2000, and best paper awards at IEEE/IFIP MMNS 2002, ACM SIGSPATIAL IWCTS 2013, and the Internet Technical Committee (ITC) 2015. He also won several other awards including best poster at IEEE INFOCOM 2017, best mobile app at ACM MobiCom 2014, and Epilepsy Foundation innovation award 2014. In 2007, 2008, and 2012, he was a finalist and winner in the ACM MobiCom SRC. In 2010, he was a winner in the ACM MobiCom WiNTech demo competition. He received teaching and mentoring recognition and awards from USC and UF. He served an editor of the Ad Hoc Networks Journal Elsevier, IEEE Computer, and IEEE Transactions on Mobile Computing, was the workshop coordination chair for ACM SIGMOBILE, and served on numerous technical IEEE and ACM committees (including chairing). His research interests include design, analysis and measurement of wireless ad hoc, sensor and mobile social networks, mobility modelling, multicast protocols, IP mobility and network simulation. He is a senior member of the IEEE and a Distinguished Scientist of the ACM.