An Information-theoretic Framework for Similarity-based, Opportunistic Social Networks

Mai ElSherief*, Tamer ElBatt†◇, Ahmed Zahran†◇, Ahmed Helmy‡

*Department of Computer Science, University of California, Santa Barbara, USA.

†Wireless Intelligent Networks Center (WINC), Nile University, Giza, Egypt.

◇Faculty of Engineering, Cairo University, Giza, Egypt.

‡The Department of Computer and Information Science and Engineering, University of Florida, Gainesville, USA.

***Abstract*—In this paper we study similarity-based networks as a key enabler for innovative applications hinging on opportunistic mobile encounters. In particular, we quantify the, inherently qualitative, notion of user similarity and introduce a novel information-theoretic framework to establish fundamental limits and quantify performance of knowledge sharing policies. First, we introduce generalized, non-temporal and temporal profile structures, beyond mere geographic location, in the form of a probability distribution function. Second, we analyze classic and information-theoretic similarity metrics using publicly available data. The most noticeable insight is that temporal metrics yield, on the average, lower similarity indices, compared to the non-temporal ones, due to incorporating the dynamics in the temporal dimension. Third, we introduce a novel mathematical framework that establishes fundamental limits for knowledge sharing among similar opportunistic users. Finally, we present numerical results characterizing the knowledge capacity for a user and the cumulative knowledge gain over time, using publicly available data for the user behavior and mobility traces, in 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, diverse scenarios and use cases as well as future knowledge sharing policies.**

Keywords: opportunistic, social networks, similarity, modeling, information theory, user traces.

I. Introduction

Recent studies, e.g. [1], point out a significant increase in the number of mobile subscribers, approaching 7 billion worldwide. This surge in mobile devices, complemented by a plethora of wireless standards, ubiquitous coverage and new use cases, has inspired novel networking paradigms and services ranging from social [2], [3] to business. However, fully understanding and exploiting the social structure of mobile users remains a daunting challenge hindering network optimization and new services. Earlier social studies, e.g., Homophily [Lazarsfeld and Merton (1954)], have shown that people tend to have similarities with others in close proximity. In such clustered communities of interest, people tend to communicate, interact and trust each other [4]. Hence, smartphones can further enrich the mobile user experience via highly personalized applications, e.g., location-based services, targeted advertisement and social networking applications among many others.

The development of similarity-based, opportunistic social networking applications would typically involve the design of three core components, namely building mobile user profiles, similarity assessment and knowledge sharing, if 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 deemed similar, the envisioned smartphone application exchanges tips about ratings, offers and other relevant information. Despite the fact that establishing trust [5], opportunistically, 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 all users trust each other and focus on candidate profile structures, user similarity, and fundamental limits of knowledge sharing.

Mobile user profiles proposed in the literature can be grouped based on different perspectives. Few are based on user location, e.g., [4], [6], while others extend the profile to capture facets beyond location, e.g., [7], [8]. From another perspective, profiles may be classified into temporal and non-temporal 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 [9]. Distance metrics from probability theory, e.g., Hellinger distance [10], can be leveraged to assess similarity between probability distribution profiles, like the ones proposed here. On the contrary, very few metrics are introduced for temporal profiles, e.g., singular value decomposition (SVD) based metrics [4].

In [11], the authors study the problem of content dissemination in opportunistic social networks. Their main result shows that high contact rate, non-social nodes (rarely found in "temporal communities") are mostly responsible for efficient content dissemination. However, mobile user profiles, similarity and the novel concepts of knowledge capacity and gain proposed here are not addressed in [11].

Our main contribution in this paper is a novel information-theoretic framework for similarity-based, opportunistic social networks. First, we extend mobile user profiles, beyond mere location, to a generalized probability distribution function and study non-temporal and temporal versions. Second, we distill key insights about known and proposed temporal and non-temporal similarity metrics, using publicly available data [12]. Third, we show the potential of the Hellinger distance to assess similarity between probability distribution user profiles and propose a novel temporal similarity metric, based on matrix vectorization, to capitalize on the richness in the temporal dimension while relying on lightweight computations. Fourth, we introduce the new notions of *Knowledge Capacity* and *Knowledge Gain*, as key quantities for formally studying knowledge sharing policies. Finally, we establish fundamental limits with the aid of information theory and unveil key insights for diverse network topologies, sharing policies and mobility scenarios and validate our theoretical findings using publicly available user behavior and mobility traces.

The rest of this paper is organized as follows. We first motivate the vision and proposed framework in Section 2. In Section 3, we study mobile user similarity with emphasis on probability distribution profiles, classic and novel metrics. In Section 4, we shift our attention to the novel information-theoretic framework to establish fundamental limits and quantify the performance of candidate knowledge sharing policies. We present key results based on realistic user mobility traces [13], [14] augmented with behavior traces from another data set [12]. Finally, conclusions are drawn and potential directions for future research are pointed out in Section 5.

II. Motivation

The wide proliferation of resource-rich smartphones renders them tightly coupled to their users, bearing a wealth of behavioral data, e.g., locations, social networks, online shopping, etc., inferring information about the user's preferences and interests. Thus, there has been growing interest in leveraging this data to open new frontiers and enrich the user's life experiences [15]. An instance of this interaction also prevails in crowd sourcing applications which may affect the real-time user behavior, e.g., Waze and Google maps provide indicators for traffic congestion and road accidents which advise the mobile users to alter their routes.

Inspired by the tight coupling between smartphones and users' behaviors, we pose the following fundamental question: Can we capitalize on the wealth of knowledge and life experiences of people we encounter throughout our lives and may have common interests, yet we do not know? The proposed framework caters to this question via an envisioned class of applications, coined *opportunistic recommendation systems* (ORS), where users capitalize on others' knowledge based on their mere co-existence and backed by homophily. The utility of ORS stems from extending our classic day-to-day "physical" recommendation exchanges, from people we know and encounter throughout the day to "virtual" exchanges with users we opportunistically encounter and do not know (yet may have things in common according to homophily) and even to users we have never encountered, through the concept of knowledge sharing/forwarding.

Finally, it is worth noting that similarity-based opportunistic social networks could serve as the basis for a variety of services, e.g., trust establishment, targeted advertisement, friend finders, and location/similarity-based services. Furthermore, ORS is expected to spur a plethora of novel smartphone applications serving large public venues, e.g., museums, theme parks, shopping malls, sports events and fairgrounds.

III. Pair-wise Mobile User Similarity

Similarity assessment is a classic problem in computer science, e.g., data mining, clustering and classification [16], [17]. For instance, it has received considerable attention for recommendation systems in online social networks [18–21]. In [18], the authors propose a model to infer relationship strength based on profile similarity and interaction activity. In [19], similarity is computed based on users' ratings of items using heuristic measures such as cosine similarity and Pearson correlation. Similarity is also studied in various contexts, e.g., web users recommendation [20] and peer recommendation systems [21].

In mobile scenarios, similarity has received limited attention through exploiting the users' spatio-temporal proximity (i.e., being at the same place at the same time), e.g. [22–24]. In [24], similar users exchange ratings about touristic places they have previously visited. In [22] and [23], users can lookup who else is in proximity and depending on common interests may decide to communicate. To the best of the authors' knowledge, similarity of mobile users has been only investigated in [22], [23], [25], [26]. However, the adopted user profile is solely based on location.

A. Generalized Mobile User Profiles

In this section, we introduce a profile structure for mobile users, beyond mere location. In addition, we explore non-temporal and temporal profiles. We assume V generic life categories, e.g., arts, sports, shopping, among others, chosen by the profile designer based on target application(s).

Thus, the non-temporal profile is a 1xV row vector where each element, C_i, captures the percentage of time spent by the mobile user, possibly online (*Interests*) or physical site visits (*Experiences*), in category i [27]. This vector can be viewed as a probability mass function (PMF) of the user profile random variable 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 in Section 4.

Inspired by [4], which proposed a temporal profile matrix for the user WiFi Access Point connectivity pattern, and the key observation that simple vector profiles hide important details about the user dynamics over time [27], we introduce probability distribution temporal profiles that capture facets other than location. Accordingly, profile vectors are captured over K time windows where each window could be a day, week, etc. depending on the dynamics of the user behavior and the time horizon. This yields a KxV profile matrix where the K profile vectors are the rows of this matrix. Deciding the time granularity and horizon, K, is a key research issue which involves data mining and analysis techniques on real-life traces capturing the users' dynamics over time and, hence, lies out of the scope of this work. For our comparative analysis, we rely on real smartphone traces from LiveLab [12] where the window is taken to be one day and K = 197 days on the average.

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

B. Similarity Metrics

The choice of similarity metrics is highly dependent on the profile representation. For non-temporal profiles, cosine and Pearson correlation are widely used in the literature [9] taking values in the ranges $[0, 1]$ and $[-1, 1]$, respectively. These metrics are widely used due to their simplicity.

Inspired by the probability distribution structure of the proposed profiles, we examine distance metrics from probability theory, namely Hellinger distance, Canberra distance and Jensen Shannon Divergence [10]. The Hellinger distance is defined for two PMFs, A and B, as [10]

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

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

On the other hand, the Canberra distance and Jensen Shannon Divergence turned out to be problematic in our context due to the fact that these two metrics yield infinite distances if there is one or more categories that are zero-valued. This is typical in practice and was frequently encountered in real-life traces, e.g., [12], where the users interests are clustered only in few categories.

On the other hand, temporal profiles lend themselves to two metrics. First, a metric based on Singular Value Decomposition (SVD) from linear algebra proposed in [4]. 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

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

which is essentially the weighted cosine similarity between the two sets of eigen-behavior vectors, where V_{Xi} and V_{Yj} are the ith and jth column of matrices V_X and V_Y, respectively. V_X and V_Y are the matrices obtained from the singular value decomposition (SVD) transformation [28] of 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 “profile anonymization”, that is, the users exchange only the elements of Σ and V, but not matrix U. This, in turn, prevents eavesdroppers from reconstructing the sender profile. On the down side, SVD similarity exhibits high computational complexity (scales quadratically with the history length K, for a fixed number of categories, V). Furthermore, similarity with oneself, $Sim_{SVD}(X,X)$, is 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 transform the two KxV profile matrices, in question, to two 1x$K.V$ vectors via the vectorization operation in Linear Algebra [28]. Afterwards, we examine classic cosine similarity.

C. Similarity Metrics Performance Comparison

In this section, we compare the performance of different similarity metrics using a real data set from the LiveLab Project at Rice University [12]. This set offers traces for smart-

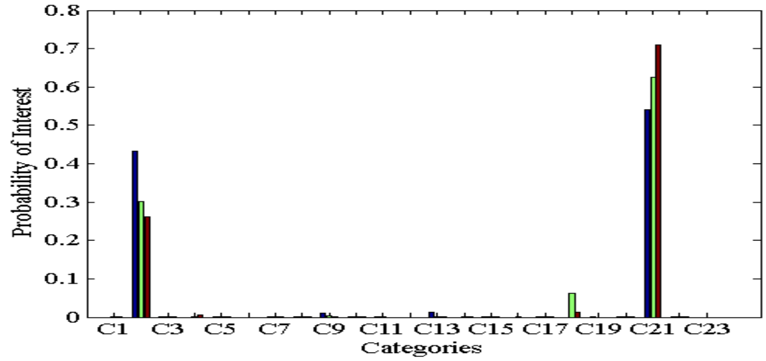

Fig. 1. Sample profile PMFs for three users from LiveLab smartphone traces.

phone users and WiFi access points (APs) from 34 iPhone 3GS users, including 24 Rice University students from Feb. 2010 to Feb. 2011 and 10 Houston Community College students from Sep. 2010 to Feb. 2011. The relevant data is stored in two database tables. The first hosts the names and genre (category) of 2500 iphone apps, as defined by Apple Store. These apps are grouped to only 23 interest categories, e.g., books, business, sports, travel, etc. The LiveLab data is particularly chosen as it readily captures categorized smartphone digital footprint logs for the mobile users as opposed to other traces in the literature which include only WiFi AP connectivity traces that are not relevant to our study. The second table includes the app usage history log for the 34 users with the date and duration of access. Cross referencing these two tables, we can generate non-temporal and temporal profiles for each user.

A remarkable observation on the distilled PMFs is that the majority of the categories in most profiles are $0-$valued and the user activity is concentrated in 2 to 5 categories as depicted in Fig. 1 and witnessed in real-life. This renders the LiveLab users "qualitatively" highly similar. This key finding hinders the use of some metrics, such as Canberra Distance and Jensen Shannon Divergence, with such sparse profiles due to the aforementioned infinite distance problem.

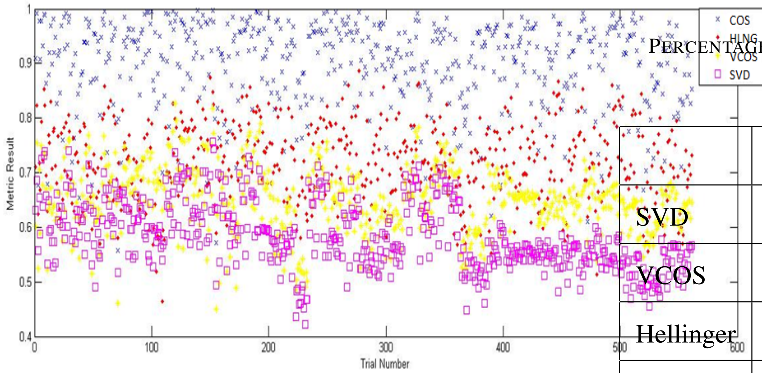

Fig. 2. Metric indices for pair-wise similarity between LiveLab users.

TABLE I

PERCENTAGE OF SIMILAR USERS FOR ALL METRICS FOR DIFFERENT SIMILARITY THRESHOLDS (T).

	T =0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
SVD	100	100	100	100	92.51	34.41	3.57	0
VCOS	100	100	100	100	98.93	80.93	18.36	0.35
Hellinger	100	100	100	100	99.82	92.87	61.5	13.3
Cos	100	100	100	100	100	99.47	97.33	89.1

Thus, we examine the cosine, Hellinger (HLNG), SVD and Vectorized cosine (VCOS) similarity metrics[1] to evaluate the pair-wise similarity for the 34 LiveLab users, which yields 561 experiments. The outcomes of the four metrics are shown, vs the experiment index, in the scatter plot depicted in Fig. 2. First, it can be noticed from Fig. 2 that cosine and Hellinger similarity yield relatively higher metric values compared to the VCOS and SVD for the same pair of users. This interesting result confirms the intuition that temporal metrics are generally "more conservative" in assessing mobile user similarity. Thus, for a given threshold T between 0 and 1, two users may be perceived "similar" using a non-temporal metric, yet, are deemed "dissimilar" using a temporal metric. This is attributed to the fact that the temporal profiles are generally more thorough since they naturally bear more details and dynamics than the non-temporal ones. This result is shown quantitatively in Table I. The table shows that VCOS and SVD yield a lower percentage of similar users than the cosine and Hellinger metrics, hence, they are more conservative. Second, the Hellinger metric may be perceived as a balance between both paradigms, since it is found to be closest to the average of the four metrics [27]. Although this demonstrates the potential of Hellinger similarity, it still deserves attention and analysis in future studies under diverse scenarios. Finally, the metrics studied and proposed here and the insights distilled open room for characterizing "actual" similarity, to serve as the ground truth in future work.

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

Based on the above observations, we envision two similarity assessment paradigms, namely macroscopic (non-temporal based) and microscopic (temporal-based), which can serve as building blocks for two-stage similarity assessment.

Macroscopic assessment: quantifies similarity between two vector-based, non-temporal profiles. Evidently, it is faster, with low-computational burden, yet, somewhat loose. Hence, it is compelling as a "coarse-grained" similarity filter.

Microscopic assessment: scrutinizes similarity between two matrix-based temporal profiles. Unlike the first paradigm, it is more conservative in declaring similarity at the expense of more complexity and, hence, being slower. It serves as a "fine-grained" similarity filter.

In the next section, we shift our attention to knowledge sharing in opportunistic encounters.

IV. Knowledge Sharing in Opportunistic Social Networks

In the rest of the paper, we shift our attention to knowledge sharing between similar users. Our prime focus is to introduce a novel information-theoretic mathematical framework, establish fundamental limits, as opposed to designing and implementing specific knowledge sharing schemes, which constitute an interesting topic of future research. This framework lays the basis for assessing the merits of future knowledge sharing and delay-tolerant forwarding policies in opportunistic social networks.

In particular, we characterize, with the aid of information theory, the amount of knowledge a user can extract in an opportunistic encounter, coined knowledge gain (KG), and the maximum amount of knowledge available for a user in the network, coined knowledge capacity (KC). The use of modeling techniques to examine the behavior of social networks is not new. For instance, graph theory has been employed extensively to gain key insights about the behavior of social networks, e.g., [29], [30] and [31]. However, employing information-theoretic tools to model the knowledge gain in opportunistic, mobile social networks has not been explored before, to the best of the authors' knowledge.

A. Opportunistic Network Model and Assumptions

The notion of a "network" here, that is, nodes exchanging information, is established solely based on pair-wise similarity, according to Section 3. 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 from now on a group of nodes where all nodes are pair-wise similar.

We model an opportunistic encounter of M similar users as a wireless adhoc network. We assume that each user is 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, modeled as a probability distribution across different categories as described in Section 3. Each user is assumed to have *tips* for sharing with similar users, e.g., recommendations of an upcoming event, best seller, site visits, etc. We assume that the users leverage short-range wireless communication technologies with fixed transmission power (i.e. the circular disk model), e.g., WiFi or Bluetooth, and, hence, interference and medium access issues are resolved using these protocols.

In this section, we wish to address two fundamental questions pertaining to knowledge sharing:

1. For an arbitrary user i, what is the maximum amount of knowledge available for this user (capacity) in a given similarity-based opportunistic encounter (network)?

2. For user i, what is amount of knowledge gain that is achievable, i.e. the user can actually reap from similar users in the network, using a specific knowledge sharing policy?

Towards this objective, we introduce next terminology and mathematical definitions.

B. Knowledge Capacity and Knowledge Gain

In this section, we introduce two new concepts that are fundamental to the analysis that follows, namely the knowledge capacity and knowledge gain.

Definition IV.1. The Knowledge Capacity(KC_i) is defined, for an arbitrary user i, as the maximum amount of knowledge that is available for user i to extract from similar users in a given network.

Definition IV.2. The Knowledge Gain (KG_i) is defined, for an arbitrary user i, as the amount of knowledge user i can gain from similar users in a given network, using a specific knowledge sharing policy.

It is straightforward to notice that $KG_i \leq KC_i$ since the knowledge capacity constitutes the upper bound on the knowledge that can be reaped out of the network, irrespective of the sharing policy. Inspired by the probability distribution definition of the user profile, we argue that probability- and information-theoretic tools would prove useful for modeling and analyzing the system at hand.

Next, we introduce the formal definition of the knowledge gain per encounter. For modeling convenience, we assume that the user tips (typically stored in a table) follow a probability distribution similar to the user profile. This does not only facilitate the mathematical analysis but is also a reasonable assumption, since users tend to have more tips and recommendations in life categories they are more interested in.

1) The Knowledge Gain per encounter: We recall from information theory that the Entropy of a discrete-valued random variable X, denoted $H(X)$, represents a measure of the "uncertainty" which also represents the amount of information this random variable bears [32]. Given our assumption that the user recommendations/tips follow the same probability distribution as the user profile, tips can be modeled as a discrete random variable, X. Accordingly, $H(X)$ quantifies the amount of information (knowledge)[2] the user has. This model opens room for formally defining the newly introduced concepts of knowledge capacity and gain.

First, we consider a toy example of an "opportunistic encounter" that involves only two users within the wireless communication range of each other. The two users have tips probability distribution vectors, denoted X and Y. Assume users X and Y [3] opportunistically meet and are deemed similar, according to Section 3. Thus, they start exchanging informative tips. Based on simple entropy relationships, depicted in the basic Venn diagram shown in Fig. 3, we distinguish three types of "knowledge" quantities (tips) that are key to our discussion:

- Tips that user X has but Y doesn't. ($H(X)$ excluding the intersection).
- Tips that user Y has but X doesn't. ($H(Y)$ excluding the intersection).
- Tips that both users have. (the intersection).

It is straightforward to map the first type of tips, which constitute the KG for user X, to the conditional entropy, $H(Y|X)$. Similarly, the KG for user Y is given by $H(X|Y)$. $KG(X)$ can also be written as

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

where $H(X, Y)$ is the joint entropy of the two random variables representing the users tips probability distributions. Finally, the third type of tips (common to both users), characterized as the mutual information $I(X; Y)$, constitutes the "communication overhead" since it is exchanged over the air despite the fact that it does not contribute to increasing the knowledge of X or Y. This perfectly agrees with our assumption that the two users know nothing about each other, when they meet opportunistically, for the first time.

It is worth noting that the knowledge capacity for user X (or Y), in this toy example of two users, is equal to the knowledge gain.

[2]We use the terms Knowledge and Information interchangebly in this paper.

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

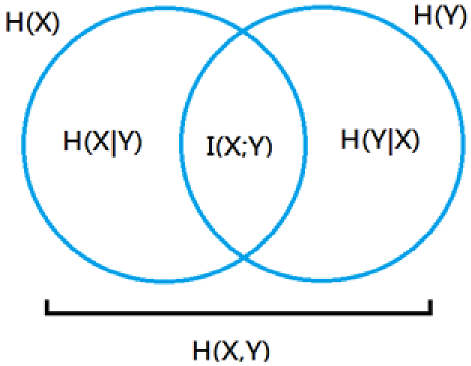

Fig. 3. Characterizing knowledge gain and communication verhead for users X and Y.

2) The Knowledge Capacity: Based on the information-theoretic definitions of the KG and KC for one encounter established in the previous section, we generalize the definition to characterize the knowledge capacity for user X_1, without loss of generality, in an opportunistic encounter with $M-1$ other users, deemed similar to X_1, as follows:

$$KC(X_1) = H(X_1, X_2, X_3, ..., X_M) - H(X_1) \quad (3)$$

which can be written as

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

(3) asserts that the maximum amount of knowledge that user X_1 can extract from the network is simply the joint information that all users have, after removing any redundant knowledge, which is represented by the joint entropy, $H(X_1, X_2, X_3, ..., X_M)$, less the amount of information that user X_1 already has, that is, $H(X_1)$. It is worth noting that (3) is valid for all network topologies and is independent of knowledge sharing policies.

This powerful mathematical framework sets the stage to establish fundamental limits and conduct analysis of diverse opportunistic social network settings under two knowledge forwarding policies. This is the subject matter of the next subsection.

C. Knowledge Sharing: Fundamental Limits and Forwarding Policies

We utilize the basic definitions introduced in the previous section to establish the KC of an arbitrary user in diverse scenarios as well as characterize its KG, under two candidate knowledge sharing policies: i) Send my tips only, or "*Send Mine Only*", whereby a user sends only his/her own tips to a similar, directly encountered user and ii) Forward my tips and others, or "*Forward Mine Plus Others*", whereby a user forwards own tips along with those acquired so far from others in previous encounters.

It is worth noting that these two policies are mere examples to illustrate the concept, however, other policies could be introduced and analyzed using the proposed model. For instance, a user could forward his/her own tips along with a subset of others' tips based on some criteria. This gives rise to a family of knowledge sharing policies that deserves a comprehensive analysis, to assess their merits and potential trade-offs, which lies out of the scope of this work.

Next, we shift our attention to quantifying the KC and KG achievable by the two aforementioned knowledge sharing policies, under a variety of opportunistic network configurations. In particular, we consider two connectivity scenarios, namely all nodes are within the communication range of each other (i.e. single-hop scenario) and multi-hop scenarios, where some nodes lie out of the communication range of others. In addition, we consider two network topology scenarios within an encounter, namely fixed topology (i.e. stationary or quasi-stationary users) and time-varying topology caused by the user's portability within the same area.

1) Fixed Topology, Similarity-based Opportunistic Networks: A. Single-hop Networks

Under this setting, the users may be stationary, quasi-stationary or even mobile, yet, each node remains one-hop away, from all other nodes, all the time. Thus, the nodes' movement does not alter the network topology, which always remains a full mesh. For this setting, we can easily characterize the knowledge capacity, as in (3), and, further, prove that the knowledge gain will achieve the capacity. This is in complete agreement with intuition since any node can take turn to exchange tips with all other nodes directly reachable. Thus, "all" knowledge that is available for any node in this network, can be reaped. The interesting question here is: how long does it take to fully acquire this knowledge? It can be easily proven that it grows linearly with the network size, M, as confirmed by simulations which reveal interesting insights discussed in Section 4.4. The achievability result for the *Send Mine Only* (SMO) policy is established by the following theorem.

Theorem 1. *For single-hop, similarity-based networks, an arbitrary node can achieve its knowledge capacity using the Send Mine Only knowledge sharing policy.*

Proof. Without loss of generality, we assume that node X_1 encounters other nodes in an increasing order of their IDs. Under the *Send Mine Only* policy, the cumulative knowledge gain for node 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 $KC(X_1)$ in (4). The same argument can be applied to all other nodes in the network which proves the result. $\square$

Corollary. If there is at least one node that is unreachable from the rest of the network and has at least one unique tip (that is not available at any other node), then $KG(X_1) < KC(X_1)$.

As indicated earlier, one of the fundamental issues in our study is how long does it take a user to achieve the knowledge capacity. This is directly related to the number of exchanges that a user needs to perform to attain the KC. Under the *Send Mine Only* policy and assuming that each node has at least one unique tip to contribute to the knowledge in the network, then it is straightforward to show that the worst-case number of exchanges needed for an arbitrary node to attain the KC is simply $(M-1)$, that is, $O(M)$.

Next, we shift our attention to quantify the KG and time-to-achievability of single-hop networks, under the *Forward Mine Plus Others* (FMPO) sharing policy. Thus, a user shares not only its own tips but also tips collected from previous encounters, denoted by the subscript p. We prove in the following theorem that the knowledge capacity is also achievable using the FMPO policy.

Theorem 2. *For single-hop similarity-based networks, an arbitrary node achieves the knowledge capacity using the FMPO knowledge sharing policy.*

Proof. We proceed along the lines of Theorem 1 and give an outline of the proof due to space limitations. Without loss of generality, we assume that node X_1 encounters all other nodes in an increasing order of their node IDs. Thus, $KG(X_1)$ based on encountering nodes $X_2, X_3, X_4, ..., X_M$ in turn, is given by

$$KG(X_1) = H(X_2, \vec{X_{2p}}|X_1) + H(X_3, \vec{X_{3p}}|X_2, \vec{X_{2p}}, X_1) + H(X_4, \vec{X_{4p}}|X_3$$
$$+ H(X_M, \vec{X_{Mp}}|X_{M-1}, \vec{X_{(M-1)p}}, ..., X_1) \tag{5}$$

where $\vec{X_{ip}}$ captures the previous encounters of node X_i. It can be shown that each joint entropy term in (5) can be expanded, e.g., $H(X_2, \vec{X_{2p}}|X_1)$ becomes $H(X_2|X_1) + H(\vec{X_{2p}}|X_2, X_1)$, where the latter term (previous encounter tips of a node) would be redundant in some cases (acquired from earlier encounters) and, hence, contributes zero to the KG. This, in turn, reduces (5) to the KC in (4) and proves the result. $\square$

It should be noted that once the conditioning, in the conditional entropy terms in the RHS, accommodates all nodes in the network, the incremental gain becomes zero and the node achieves its knowledge capacity. In essence, the role of the previous encounters (prevails in the conditional entropy terms) is the prime contributor to the FMPO policy attaining the KC faster, compared to the SMO policy. Apparently, this does not come for free since there is a fundamental trade-off between the cumulative KG after a number of exchanges and the associated overhead which warrants attention in future research, especially in multi-hop networks. We prove next that the communication overhead of FMPO is greater than or equal to SMO, in single-hop networks.

Theorem 3. *For single-hop networks, the communication overhead under FMPO is greater than or equal to SMO.*

Proof. Assume two users, X, Y, encounter each other. Denote the vector of previous encounters for X and Y by $\vec{X_p}$ and $\vec{Y_p}$, respectively.

Under the SMO policy, the communication overhead that X incurs is the common knowledge (mutual information) between what X sends (which is its knowledge only) and Y's knowledge so far which is given by

$$OH(X)_{SMO} = I(X; Y, \vec{Y_p}). \tag{6}$$

Similarly, the communication overhead from the perspective of user Y is $OH(Y)_{SMO} = I(Y; X, \vec{X_p})$.

Under FMPO, the communication overhead is the same for both users, X and Y, and is given by

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

From information theory, the mutual information between two random variables A and B can be written as

$$I(A; B) = H(A) + H(B) - H(A, B). \tag{8}$$

Applying (8) on (6) yields

$$OH(X)_{SMO} = H(X) + H(Y, \vec{Y_p}) - H(X, Y, \vec{Y_p}). \tag{9}$$

Applying (8) on (7) yields

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

Subtracting (9) from (10) yields

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

Since $H(A, B) = H(A) + H(B|A) = H(B) + H(A|B)$, (11) could be re-written as

$$OH(X)_{FMPO} - OH(X)_{SMO} = H(X) + H(\vec{X_p}|X) - H(X) - [H(\vec{X_p}|X,$$

which reduces to

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

$$\geq 0 \tag{13}$$

where the inequality in (13) follows since conditioning reduces entropy. This proves the result. $\square$

B. Fixed Topology Multi-hop Networks

Under this setting, the network topology is time-invariant, however, not all nodes are directly reachable from each other, i.e. multi-hop paths will be followed for exchanging some tips.

In this case, the role of the knowledge sharing policy stands out and affects whether a user can/cannot attain the KC.

Next, we quantify the performance, and trade-offs, of the two aforementioned knowledge sharing policies, namely SMO and FMPO. In case of SMO, the knowledge gain achieved by node X_1 is limited by the neighborhood size, $N<M$. This renders the KG under this policy strictly less than the KC for multi-hop networks. The following theorem formally establishes this result.

Theorem 4. *For fixed topology, multi-hop, similarity-based networks, the Send Mine Only knowledge sharing policy will not attain capacity, that is, $KG(X_1) < KC(X_1)$ iff $N < M$.*

Proof. Without loss of generality, we assume node X_1 communicates with other nodes in an increasing order of their IDs. The cumulative knowledge gain for node X_1, $KG(X_1)$, according to exchanges with neighbors $X_2, X_3, X_4, ..., X_N$ is given by $H(X_2|X_1)+H(X_3|X_2, X_1)+.....+H(X_N|X_{N-1}, ..., X_1)$. It is worth noting that the summation of positive conditional entropy terms is limited to $N < M$ nodes. It misses other positive terms involving the $M - N$ non-neighbors to X_1. Hence, it directly follows that $KG(X_1) < KC(X_1)$, which proves the result. □

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 4.3.1.A that the KC 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 capacity, even if $N < M$.

In the rest of this section, we focus on the performance of the FMPO knowledge sharing policy for fixed topology multi-hop networks. As expected, forwarding others tips opens room for a node to achieve its KC, even when $N < M$. The following theorem formally establishes this result.

Theorem 5. *For fixed topology, multi-hop, similarity-based networks, an arbitrary node can achieve the knowledge capacity using the FMPO knowledge sharing policy.*

Proof. Without loss of generality, we assume that each node starts off with its own knowledge only and X_1 encounters other nodes in an increasing order of their IDs, while other neighbors have pair-wise encounters with other nodes in the network. The cumulative knowledge gain for node X_1, $KG(X_1)$, after communicating with neighboring nodes $X_2, X_3, X_4, ..., X_N$ is given by

$$KG(X_1) = H(X_2, |X_1) + H(X_3, \vec{X_{3p}}|X_2, X_1) + H(X_4, \vec{X_{4p}}|X_3, \vec{X_{3p}}, X_2$$

$$+H(X_N, \vec{X_{Np}}|X_{N-1}, \vec{X_{(N-1)p}}, ..., X_1) \quad (14)$$

where $\vec{X_{ip}}$ represents the previous encounters of node i. At this point, two cases arise. First, if the previous encounters include "forwarded" tips from all non-neighboring nodes, namely $X_{N+1}, X_{N+2}, ..., X_M$, then it can be shown that the cumulative knowledge gain of X_1 is given by

$$KG(X_1) = H(X_1, X_2, X_3, ..., X_M) - H(X_1) = KC(X_1) \quad (15)$$

which attains the $KC(X_1)$. On the other hand, if the previous encounters do not cover tips from non-neighbors, then this implies that node X_1 needs more time to attain $KC(X_1)$, possibly via revisiting neighbors that it has already visited to acquire the missing knowledge from nodes out of its communication range, until it eventually reaches capacity. This proves the theorem. □

D. User Traces and Performance Results

In this section, we back our theoretical findings with numerical results based on smartphone traces for user profiles [12] and real-life user mobility traces, gathered at Infocom 2005 [13], [14].

1) Single-hop Similarity-based Networks: In this section, we rely on real-life traces, either for user behavior or mobility. For user behavior, we utilize digital footprint traces for 20 smartphone users from the LiveLab project [12]. In this section, the users are assumed stationary, or quasi-stationary, and, hence, the network topology is a time-invariant full mesh. In order to quantify the knowledge capacity and gain for an arbitrary user in this network, we need to pre-process the huge amount of data in two steps. First, we compute the joint probability mass function for the 20 users under investigation over a period of six months, from September 2010 to February 2011. Thus, we monitor the users' activities categorized under 24 categories[4], each second, for six months and record their concurrent activities. Afterwards, we divide by the total duration of the six months to get the joint PMF. Next, we show the cumulative knowledge gain increase as the node under investigation encounters more nodes over time.

First, we present the performance results for a single-hop network under the SMO knowledge sharing policy. For the network of $M = 20$ nodes discussed earlier, an arbitrary user can achieve the knowledge capacity in $M-1 = 19$ exchanges. This is shown in Fig. 4 for three arbitrary users, namely $B00$, $B04$ and $D03$. It is also worth noting that the cumulative knowledge gain is a non-decreasing function with time. On the other hand, the knowledge capacity is shown as a horizontal solid line that is generally different from one node to another.

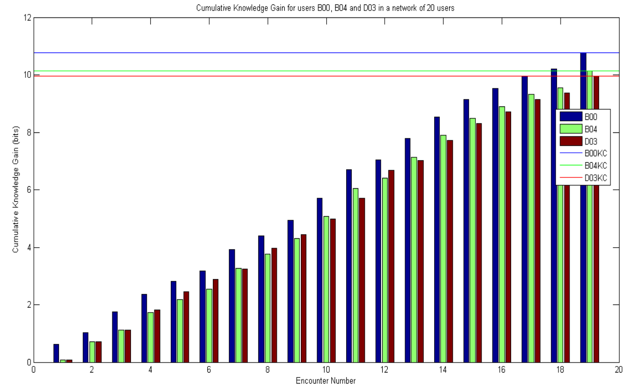

Fig. 4. Cumulative knowledge gain for three users in a fixed topology single-hop network under SMO.

Next, we shift our attention to the same network setting, yet, employing the FMPO policy. Based on Theorem 2, we show that, for this type of networks, all nodes achieve the knowledge capacity using the FMPO policy, yet, faster than SMO, i.e. in less encounters due to sharing the tips of others. This valuable insight is confirmed for users $B00$, $B04$ and $D03$ in Fig. 5.

2) Fixed Topology, Multi-hop Similarity-based Networks: In this section, we shift our attention to fixed topology, multi-hop networks where we study the cumulative knowledge

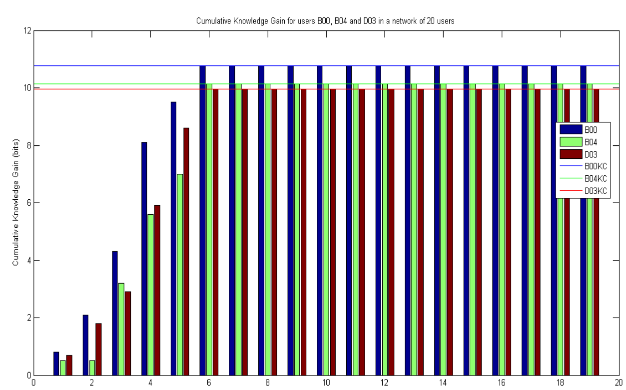

Fig. 5. Cumulative knowledge gain for three users in a fixed topology single-hop network under FMPO.

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

gain behavior under the SMO and FMPO knowledge sharing policies.

As indicated earlier, and shown formally in Theorem 4, achieving the KC of an arbitrary node using SMO is fundamentally limited by the single-hop neighborhood size of this node, denoted N. For instance, we generate 20 randomly generated topologies of uniformly distributed users whereby each user has 6-7 single-hop neighbors, out of 20 nodes, on the average. Thus, $B00$ does not achieve the KC for it in this network, as established in Theorem 4 and shown here using real smartphone user behavior traces in Fig. 6. Thus, the maximum KG node $B00$ can achieve is only 43% of its KC. Similarly, users $B06$ and $D00$ have single-hop neighbors strictly less than $M = 20$ and, hence, cannot achieve their respective KCs.

In the rest of this section, we analyze the KG and KC 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' knowledge (tips) and, hence, the nodes could achieve the knowledge capacity as proven in Theorem 5. The results here are based on the 20 randomly generated topologies discussed earlier. The cumulative knowledge gains for users $B00$, $B06$ and $D00$ are depicted in Fig. 7. We notice that the KC is achievable for the three shown users after 8, 9 and 10 encounters, respectively.

5) Time-varying Topology (Mobile) Similarity-based Networks: A. User Profiles and Mobility Traces

After an extensive search for mobile user traces on publicly available data repositories, e.g., CRAWDAD [13], [14] and the alike, we did not find traces that include both, user behavior and mobility patterns. Moreover, most of the mobility traces are university campus WiFi access patterns as opposed to user encounters. In order to proceed with the performance evaluation based on real data, we resort to jointly leveraging traces from different data sets, for the profiles and mobility. As for user profiles, they are distilled and constructed based on the LiveLab project data [12] described earlier. On the other hand, user mobility traces are based on a "conference encounter" data, namely Infocom 2005 [13], [14]. Towards this objective, we project the user profile probability distributions, constructed from the LiveLab data, on the Infocom user mobility traces. For the Infocom 2005 experiment, the data set is relatively small whereby participants are 50 attending

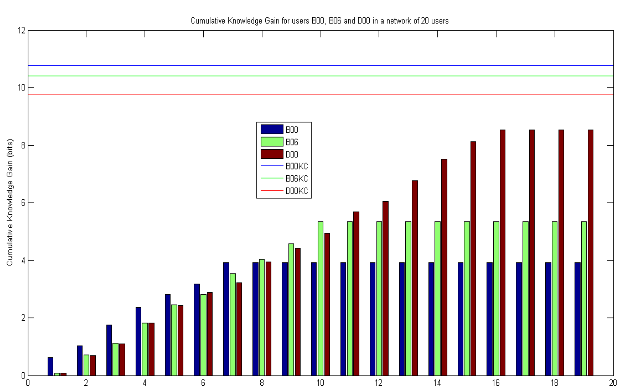

Fig. 6. Cumulative knowledge gain for three users in a fixed topology multi-hop network under SMO.

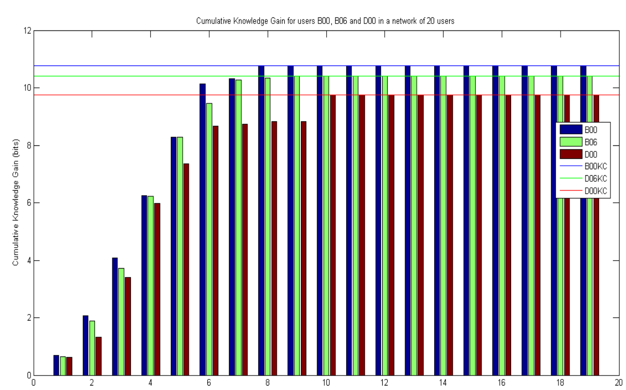

Fig. 7. Cumulative knowledge gain for three users in a fixed topology multi-hop network under FMPO.

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. Time-varying Topology Multi-hop Networks

In this section, we quantify the knowledge capacity 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 KC and, if so, how much time this would incur. The gathered results are shown in Table II. 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 to approach the knowledge capacity. On the other hand, some nodes, e.g., $B06$, do not benefit from mobility since they remain disconnected throughout the experiment time span. This insight agrees with intuition since the mobility patterns of some nodes could assist them in encountering the "knowledge hot spots" of the network. 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 no less KG 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., [14], [33], 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 theorem, we establish the proof based on an ideal mobility model guaranteeing encounter with all other nodes, which may not be valid for a whole campus scenario based on [33]. 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 [34], the following theorem proves that the knowledge capacity in mobile, delay-tolerant, multi-hop social networks is always achievable,

TABLE II

KNOWLEDGE GAIN FOR TEN MOBILE USERS AFTER HALF A DAY.

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

under idealistic assumptions and loose delay constraints. Those assumptions guarantee that an arbitrary node will encounter all other nodes in the network, almost surely. Nevertheless, characterizing the realistic mobility scenarios and conditions under which the KG of a node is improved by mobility is an interesting subject of future research.

Theorem 6. *For a time-varying topology, multi-hop, similarity-based network, an arbitrary node achieves its knowledge capacity, almost surely, in case each node moves according to an independent 2-dimensional random walk in a fixed area and under loose delay constraints.*

Proof. In case of loose delay constraints and independent 2-dimensional random walks, an arbitrary node will encounter all other nodes in the network, almost surely (follows from Lemma 6 in [34]). Hence, without loss of generality, we assume that node X_1 encounters all nodes in an increasing order of their node IDs. Under 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 theorem. □

V. CONCLUSION

In this paper we propose a novel information-theoretic framework for similarity-based opportunistic social networks. We first introduce generalized, non-temporal and temporal profiles in the form of a probability distribution function. Second, we analyze classic and information-theoretic similarity metrics using publicly available data. The most noticeable insight is that temporal metrics yield, on the average, lower similarity indices, compared to the non-temporal ones, due to incorporating the dynamics in the temporal dimension. Third, we introduce a novel mathematical framework that establishes fundamental limits and insightful results for two sample knowledge sharing policies among similar opportunistic users. Finally, we present numerical results characterizing the knowledge capacity for a user and the cumulative knowledge gain over time, using publicly available traces for user behavior and mobility, in case of fixed and mobile scenarios. This work can be extended along multiple directions, e.g., novel similarity metrics capitalizing on the strengths of non-temporal and temporal profiles, examine Hellinger and vectorized cosine similarity with diverse users and scenarios, leverage the proposed mathematical framework to analyze novel knowledge sharing policies and, finally, establish fundamental limits for mobile scenarios.

REFERENCES

[1] ITU, "ITU releases 2014 ICT figures," http://www.itu.int/net/pressoffice/press_releases/2014/23.aspx/, accessed: 02/19/2015.

[2] I. Smith, "Social-mobile applications," *Computer*, vol. 38, no. 4, pp. 84–85, 2005.

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

[4] W. Hsu, D. Dutta, and A. Helmy, "CSI: A paradigm for behavior-oriented profile-cast services in mobile networks," *Ad Hoc Networks*, vol. 10, pp. 1586–1602, 2012.

[5] S. Trifunovic, F. Legendre, and C. Anastasiades, "Social trust in opportunistic networks," in *IEEE INFOCOM Worshop.*, 2010, pp. 1–6.

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

[7] J. Ramer, A. Soroca, and D. Doughty, "Mobile user profile creation based on user browse behaviors," Oct. 30 2007, US Patent App. 11/929,129.

[8] S. Moghaddam, A. Helmy, S. Ranka, and 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*, 2010, pp. 248–256.

[9] W. Jones and G. Furnas, "Pictures of relevance: A geometric analysis of similarity measures," *Journal of the American Society for Information Science*, vol. 36, pp. 420–442, 1987.

[10] S. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *City*, vol. 1, no. 2, p. 1, 2007.

[11] A.-K. PietilÃd'nen and C. Diot, "Dissemination in opportunistic social networks: the role of temporal communities," in *ACM Mobihoc*, 2012.

[12] C. Shepard, A. Rahmati, C. Tossell, L. Zhong, and P. Kortum, "Livelab: measuring wireless networks and smartphone users in the field," *ACM SIGMETRICS Performance Evaluation Review*, vol. 38, no. 3, pp. 15–20, 2011.

[13] "CRAWDAD trace set cambridge/haggle/imote (v. 2009-05-29)," Downloaded from http://crawdad.cs.dartmouth.edu/cambridge/haggle/imote.

[14] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, 2007.

[15] N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 36, pp. 15 274–15 278, 2009.

[16] R. T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in *Proc. of*, 1994, pp. 144–155.

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

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

[19] I. Konstas, V. Stathopoulos, and 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*, 2009, pp. 195–202.

[20] M. Pazzani, "A framework for collaborative, content-based and demographic filtering," *Artificial Intelligence Review*, vol. 13, no. 5, pp. 393–408, 1999.

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

[22] S. 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.

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

[24] A. De Spindler, M. Norrie, M. Grossniklaus, and B. Signer, "Spatio-temporal proximity as a basis for collaborative filtering in mobile environments," *Proceedings of UMICS*, pp. 912–926, 2006.

[25] M.-J. Lee and 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.

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

[27] M. ElSherief, T. ElBatt, A. Zahran, and A. Helmy, "The quest for user similarity in mobile societies," in *The 2nd International Worskhop on Social and Community Intelligence, in conjunction with IEEE Percom.*, 2014.

[28] G. Strang, "The fundamental theorem of linear algebra," *American Mathematical Monthly*, pp. 848–855, 1993.

[29] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1979.

[30] M. E. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. Suppl 1, pp. 2566–2572, 2002.

[31] J. A. Barnes, "Graph theory and social networks: A technical comment on connectedness and connectivity," *Sociology*, vol. 3, no. 2, pp. 215–232, 1969.

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

[33] W.-j. Hsu and A. Helmy, “On nodal encounter patterns in wireless lan traces,” *IEEE Transactions on Mobile Computing*, vol. 9, no. 11, pp. 1563–1577, 2010.

[34] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah, “Throughput-delay trade-off in wireless networks,” in *IEEE INFOCOM’04*, 2004.