Modeling and Estimating the Structure of D2D-Based Mobile Social Networks

Sigit Aryo Pambudi Wenye Wang
Department of Electrical and Computer Engineering
North Carolina State University, Raleigh, NC 27606
Email: {sapambud,wwang}@ncsu.edu

Cliff Wang
Army Research Office
Research Triangle Park, NC 27709
Email: cliff.wang@us.army.mil

***Abstract*—Along with the explosive growth of mobile social network (MSN) users and the advent of device-to-device (D2D) communications, D2D-based MSN (D2D-MSN) has become a promising alternative for exchanging multimedia contents on-the-go. Although the complete structure of a D2D-MSN plays a key role in understanding its performance, such knowledge is not readily available due to the difficulty of collecting connectivity information from the vast amount of users. To model the structure, we define a D2D-MSN network that jointly captures the social connectivity over the MSN and the opportunistic D2D contacts among users. A random walk with self loop (RWSL) scheme that quickly converges to its stationary distribution is proposed to collect a subset of D2D-MSN nodes. An estimator is then introduced to obtain an unbiased estimate of the D2D-MSN graph's joint degree distribution, p_{ij}, from the set of visited nodes, leading to an unbiased RWSL scheme. The resulting estimate of p_{ij} can be used as a statistic for creating synthetic graph and generating functions for analyzing robustness of D2D-MSN. Numerical results show that the proposed unbiased RWSL converges faster to its stationary distribution, achieves higher joint degree distribution accuracy, and visits less number of nodes, compared to existing graph exploration schemes.**

I. Introduction

The rapid development of mobile communication technologies has resulted in an explosive growth of mobile internet users. Among the internet services accessed on-the-go, online social networking services (SNSs) like Facebook, Twitter, and Google+ have become one of the most popular, accessed by as many as one out of four people worldwide [1], changing how information spreads from a "word-of-mouth" paradigm into a "word-of-text, -audio, and -video" fashion. This, in combination with successful miniaturization of mobile devices have motivated vast amount of users to access SNSs through their smartphones and/or tablets, leading to a mobile social network (MSN) paradigm. Recently, the concept of device-to-device (D2D) networking [2], [3] that exploits opportunistic short-range contact through Bluetooth, WiFi Direct, and near-field communication (NFC) in order to offload traffic from backbone networks has emerged as a promising alternative to MSN accessed over centralized networks [4], introducing a new paradigm called D2D-based MSN (D2D-MSN). For the vast mobile SNS users, whose number is projected to grow to 3.1 billion in 2018 [5], this new paradigm offers benefit such as prolonged battery life due to reduced transmission power and higher data rate due to shorter communication range [2].

This work is supported in part by Army Research Office under grant number W911NF-15-2-0102 and NSF Award number CNS 1423151.

Motivated by the vast number of potential mobile users that may depend their information sharing activities on D2D-MSN, there exists a fundamental need to understand the quality of information dissemination in such a new paradigm. In terms of information dissemination, gossip spreading and viral marketing [6] are among the most studied applications, whose performance can readily be evaluated when the complete structure of the social network under examination is available at hand. The complete structure of a D2D-MSN, however, is not known *a priori* since the number of users can be very large in practice such that collecting the local connectivity information from all the users is impractical. In the field of complex network, methods for estimating the properties of a *single* network using graph crawling and sampling have been widely proposed. In [7], graph traversals based on breadth-first search for sampling Internet topologies has been proposed, while graph explorations based on random walk and Metropolis-Hastings algorithms have been applied to Facebook social graph [8]. But, the structure of a D2D-MSN is different; It consists of an MSN network as well as a D2D network that together determine the information dissemination dynamics, such that the existing estimation schemes for single networks cannot be applied directly. To this end, two fundamental questions remain unanswered: *"How to model the structure of a D2D-MSN that comprises two coexisting MSN and D2D networks? How to estimate the structure of the two, possibly correlated MSN and D2D networks?"* The answers are critical toward understanding the information dissemination performance of a D2D-MSN that may determine the success of its application.

As mentioned above, our objective is to model and estimate the structure of a D2D-MSN. Since the source and destination of social contents in a D2D-MSN are described by the inter-user relationship on the MSN while the actual path taken by such contents is governed by the D2D network, we introduce a D2D-MSN graph $\mathcal{G}(t)$ that combines the inter-user connectivity over both the MSN (social) and D2D (communication) networks. In $\mathcal{G}(t)$, the marginal degree distributions of the social and communication graphs are well known to represent the respective graphs' structures [9] and further describe information dissemination dynamics in each graph [10], but cannot capture how the social graph $\mathcal{G}_s(t)$ is coupled to the communication graph $\mathcal{G}_m(t)$. Thus, to capture the relationship between $\mathcal{G}_s(t)$ and $\mathcal{G}_m(t)$, we consider a *joint degree distribution* p_{ij} that represents the probability a user has i and j neighbors in the social and communication graphs, respectively, and use this metric to characterize the structure of $\mathcal{G}(t)$. As a result, estimating the structure of $\mathcal{G}(t)$ becomes equivalent to the problem of estimating p_{ij}.

To take on the problem of estimating p_{ij}, we first collect a subset of nodes in $\mathcal{G}(t)$ by walking over the communication graph. To regulate the progression of such walk over time, we assign a transition probability from each visited node to its immediate neighbors, such that the time until the walk converges to its stationary distribution given local information only is minimized. This leads to a random walk with self-loop (RWSL) scheme. Using the history of nodes visited by the RWSL scheme, the joint degree distribution, p_{ij}, is then estimated as the sample mean of the visited nodes' degree distribution. Using first-order approximation, we show that the resulting estimate, p'_{ij}, is inherently *biased*, i.e. there exists a gap between the expected estimate, $\mathbb{E}(p'_{ij})$, and the target distribution, p_{ij}. To eliminate such bias, we adapt the Hansen-Hurwitz estimator [11] and propose a bias correction scheme. The combination between the bias correction and the RWSL scheme, which is referred as *unbiased RWSL* (Unb-RWSL), provide a methodology for obtaining an unbiased estimate of p_{ij} from a D2D-MSN whose structure is *a-priori* unknown.

The performance of the Unb-RWSL is then validated through numerical simulations. Numerical results show that the proposed RWSL scheme is shown to converge faster to both its stationary distribution and its actual covariance than two well-known graph exploration algorithms: Metropolis-Hastings (MH) and random walk (RW) [8]. Furthermore, the proposed Unb-RWSL scheme is shown to achieve higher joint degree distribution accuracy than the MH scheme. Finally, given the same number of iterations, the proposed Unb-RWSL visits less number distinct nodes than the RW scheme, thus requiring less amount of memory to store the visited nodes' information.

In this paper, Section II introduces the D2D-MSN graph model and the problem of estimating its joint degree distribution, p_{ij}. An Unb-RWSL scheme that achieves fast and unbiased estimation of p_{ij} is proposed in Section III and evaluated in Section IV. Finally, Section V concludes the paper.

II. Network Model and Problem Formulation

In this section, we first introduce a D2D-MSN graph model, and proceed by stating the main problem of finding an unbiased estimate of the D2D-MSN's joint degree distribution.

A. D2D-Based Mobile Social Network Graph Model

Since a D2D-MSN consists of coexisting social (MSN) and communication (D2D) networks, let us examine the MSN counterpart first. Let $\mathcal{V}$ denotes the set of $n = |\mathcal{V}|$ users in a D2D-MSN. In this paper, the terms 'user' and 'node' will be used interchangeably. Each user in $\mathcal{V}$ has an account in a "virtual" world called mobile social network (MSN). Let $\mathcal{E} = \mathcal{V} \times \mathcal{V}$ denotes the set of all possible edges between users and $t \in \mathbb{N}_0$ be a discrete time slot. Let $\mathcal{E}_s(t) \subset \mathcal{E}$ be the set of all undirected edges between users in the MSN at time t. Then, users u and v may exchange social contents if they have an edge in the MSN, *e.g.*, $(u, v) \in \mathcal{E}_s(t)$. Examples of such *social edge* are friendship relationship between users in Facebook and social circles in Google+. Let social neighbors $\mathcal{N}_s(u)$ of u be the set of users in $\mathcal{V}$ that has a social edge with u as its endpoint, *e.g.*, $\mathcal{N}_s(u) := \{v \in \mathcal{V} : (u, v) \in \mathcal{E}_s(t)\}$. When a user generates a content in the MSN, *i.e.*, a *micropost* [4], it will forward such content to its social neighbors. Thus, $\mathcal{E}_s(t)$ characterizes all the possible source-destination pairs for social contents in an MSN. Put together, $\mathcal{E}_s(t)$ and $\mathcal{V}$ form the tuple $\mathcal{G}_s(t) = (\mathcal{V}, \mathcal{E}_s(t))$, which is referred as a *social graph*.

In the D2D-MSN, users access the MSN on-the-go through their respective personal devices, such as smartphone and tablet. Through the MSN, the users interact with each other by exchanging social contents, which is physically sent as wireless packets called messages. We assume that D2D wireless communication, such as Bluetooth and NFC, is used to exchange messages between users and the transmission range is finite due to limited battery power and health concerns. Let $\mathcal{E}_m(t) \subset \mathcal{E}$ be the set of bi-directional edges between nearby users that can exchange messages at time slot $t \in \mathbb{N}_0$, called *communication edges*. We assume that *multihop* communication is enabled, *i.e.*, two users $u \in \mathcal{V}$ and $v \in \mathcal{V}$ can exchange messages if an end-to-end path represented by a sequence of communication edges $\{(u, v_1), (v_1, v_2), \ldots, (v_{k-1}, v)\}$ exists in time slot t. Here, $v_1, \ldots, v_{k-1} \in \mathcal{V}$ may not be the social neighbors of both u and v, but still participate in *relaying* messages between u and v. As a result, the set of communication edges $\mathcal{E}_m(t)$ characterizes the possible path(s) taken by messages exchanged by any two arbitrary users in an MSN. Combined with $\mathcal{V}$, $\mathcal{E}_m(t)$ forms the tuple $\mathcal{G}_m(t) = (\mathcal{V}, \mathcal{E}_m(t))$, which is referred as a *communication graph*. Since the number of new friend increases slower than the user movement rate, we assume that the communication graph changes faster than the social graph, e.g., $|\frac{d}{dt}\mathcal{E}_m(t)| > |\frac{d}{dt}\mathcal{E}_s(t)|$.

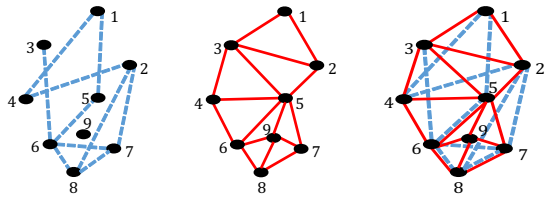
To better understand how $\mathcal{G}_s(t)$ and $\mathcal{G}_m(t)$ together form a D2D-MSN, let us examine the following illustration. Examples of social and communication graphs are respectively depicted in Figs. 1(a) and 1(b). In Fig. 1(a), users 2 and 7 may exchange social contents, but the shortest path traversed by such contents in the communication graph, $\{(2, 5), (5, 7)\}$, have the length of two. On the other hand, the shortest paths $\{(2, 5), (5, 7), (7, 8)\}$, $\{(2, 5), (5, 9), (9, 8)\}$ and $\{(2, 5), (5, 6), (6, 8)\}$ traversed by the social contents exchange between users 2 and 8 have the length of three. Note that user 9 is a relay node. From the social graph perspective, we also assume that social contents received by a user may be re-shared to its social neighbor, *i.e.*, a social content from 2 received by user 7 may be re-shared to user 8. Thus, 7 is potentially an important user, because (i) it lies in the shortest communication path of at least two pairs of social content exchange activities, and (ii) it can help re-share social contents from 2 to 8. This illustration shows that put together, the social and communication graphs provide more information regarding a user's importance toward social content dissemination process, which is useful, *e.g.*, for selecting seed node(s) for content spreading in mobile advertisements [4] and for generating the user activity graphs [12] of a D2D-MSN. Motivated by this usefulness, we combine $\mathcal{G}_s(t)$ and $\mathcal{G}_m(t)$ and denote a D2D-MSN graph at time $t \in \mathbb{N}_0$ as

$$\mathcal{G}(t) = (\mathcal{V}, \mathcal{E}_s(t), \mathcal{E}_m(t)). \tag{1}$$

A $\mathcal{G}(t)$ obtained from combining the communication and social graphs in Figs. 1(a) and 1(b) is depicted in Fig. 1(c).

B. Problem Formulation

Commercial SNS providers such as Twitter, Google+, and Facebook do not openly provide the complete information



(a) Social graph. (b) Commun. graph. (c) D2D-MSN graph.

Fig. 1. An example of a D2D-MSN.

regarding the inter-user relationship (i.e., $\mathcal{E}_s(t)$) due to privacy issues. Moreover, in the D2D network considered here, there is no centralized authority that collects the complete information regarding the user-to-user contacts, which is characterized by $\mathcal{E}_m(t)$. The number of users, n, of a D2D-MSN may also be very large such that keeping track of all the users' social and D2D contacts may not be possible even under the existence of a centralized authority. Thus, the full knowledge regarding the D2D-MSN graph $\mathcal{G}(t)$ is not readily available. Knowing the structure of $\mathcal{G}(t)$, however, is important since it will allow us to analyze and improve the performance of the examined D2D-MSN [4], [12]. Motivated by this, we are interested in estimating the structure of the D2D-MSN graph $\mathcal{G}(t)$.

Regarding our interest above, we first determine a metric that can capture the structure of the D2D-MSN graph. In the complex network literature, *degree distribution*, which denotes the number of neighbors connected to a node, is a centrality metric that has been widely used to characterize the structure of a graph [9] and to generate functions for describing message epidemic over graphs [10], from which several properties, such as percolation phase transition for stand-alone and coupled networks under attacks [13], have been analyzed. Moreover, by taking the degree distribution as its input, *configuration model* (CM) method can be employed to generate a graph whose degree converges to the input distribution, assuming erasure scheme is applied to eliminate the multiple edges and self-loops potentially created by pure CM scheme [9]. Motivated by its vast usefulness, then we are interested in estimating the node degree distribution of the D2D-MSN graph $\mathcal{G}(t)$.

With such objective at hand, we further ask, "How to quantify whether the estimated degree distribution is a good representative to that of $\mathcal{G}(t)$?" To answer this, let us re-visit the D2D-MSN graph $\mathcal{G}(t)$. We assume that an algorithm for estimating the degree distribution of $\mathcal{G}(t)$ starts at time $t \in \mathbb{N}_0$ and completes before $t+1$, while the D2D-MSN graph is stationary during $[t, t+1)$. Because $\mathcal{G}(t)$ consists of both the social and communication graphs, then it will be jointly represented by two degree distributions: one for $\mathcal{G}_s(t)$ and one for $\mathcal{G}_m(t)$. Let *joint degree distribution* $p_{ij} := Pr\{d_s(u) = i, d_m(u) = j\}$, $\forall i, j \in \{0, 1, \ldots, n-1\}$ be the probability that a randomly-selected user $u \in \mathcal{V}$ has i and j neighbors in $\mathcal{G}_s(t)$ and $\mathcal{G}_m(t)$, respectively. The joint degree distribution is known to fully characterize a two-variate random variable, precisely what the degrees of $\mathcal{G}(t)$ are. Let there be a process that generates an estimated joint degree distribution of $\hat{p}_{ij}$. Also, let p_{ij} be the actual joint degree distribution of $\mathcal{G}(t)$, referred as a *target distribution*. Then, we have the following definition.

Definition 1: $\hat{p}_{ij}$ is an unbiased estimate of p_{ij} if $\mathbb{E}(\hat{p}_{ij}) = p_{ij}$, for all i, j, where $\mathbb{E}(\cdot)$ is an ensemble mean operator.

This implies that, in average, a graph generated, *e.g.*, by the CM method, using $\hat{p}_{ij}$ will *not* differ from that using the original D2D-MSN graph $\mathcal{G}(t)$, as long as $\hat{p}_{ij}$ is unbiased to p_{ij}. For this reason, the objective of obtaining the structure of $\mathcal{G}(t)$ can be re-stated as *"to obtain an estimated joint degree distribution $\hat{p}_{ij}$ that is unbiased to the target distribution p_{ij}."*

III. Fast and Unbiased Estimation of Joint Degree Distribution

In this section, we discuss a methodology for achieving a fast, unbiased estimation of the target distribution by proposing a random walk with self-loop (RWSL) scheme for obtaining a subset of the nodes in $\mathcal{G}(t)$ in the smallest amount of time, using which an intermediate estimate of joint degree distribution, p'_{ij} is calculated. Since the employed RWSL scheme induces a bias to p'_{ij}, we then outline a bias correction scheme to to obtain an unbiased $\hat{p}_{ij}$.

A. Random Walk Over D2D-MSN

Before the target distribution can be estimated, a relatively small but representative subset of nodes, denoted as $\mathcal{V}' \subset \mathcal{V}$, must first be obtained from the set of nodes $\mathcal{V}$ in $\mathcal{G}(t)$. The main reason of doing so is because p_{ij} will be estimated using the sample mean of the obtained subset. Since the complete structure of $\mathcal{G}(t)$ is not known *a priori*, we employ a graph *walking* algorithm to obtain the subset $\mathcal{V}'$. A walker is a program that performs walking over a graph, which can be considered as a file running in a node's memory. Then, *"moving"* a walker consists of stopping the program at the current node, transmitting it as a message over the D2D network, and executing it at the next node. Let a walker starts by initially selecting a node $u \in \mathcal{V}$ at random. We assume that node u knows its immediate neighbors $\mathcal{N}_s(u)$ and $\mathcal{N}_m(u)$ in the social and mobile graphs, respectively. At every iteration, the walker (a) records $|\mathcal{N}_m(u)|$ and $|\mathcal{N}_s(u)|$; (b) assigns a transition probability $P(u, v)$ to all $v \in \mathcal{N}_m(u) \cup u$; (c) selects one node among $\{v \in \mathcal{N}_m(u) \cup u\}$ according to $P(u, v)$; (d) moves to the selected node; and then (e) re-assigns the selected node as u. The process is iterated until sufficient number of nodes have been visited by the walker. The visited nodes themselves are then used as the set $\mathcal{V}'$.

In the graph walking algorithm above, we assume that the walker may select the next node only among the communication neighbors, $\mathcal{N}_m(u)$, and node u itself. The *first* reason behind this is because there exists relay nodes that has communication neighbors but are not connected socially. A walk over the social graph will never visit these kind of nodes. On the other hand, by selecting the next nodes among the communication neighbors and u itself, such relay nodes will be taken into account into the estimation of p_{ij}. The *second* reason is because the communication graph's rate-of-change is faster than that of the social graph. Thus, performing a walk over $\mathcal{G}_m(t)$ every time the communication graph changes will sufficiently capture any change in both $\mathcal{G}_m(t)$ and $\mathcal{G}_s(t)$. Note that both of the aforementioned objectives can be achieved if the walk is performed over the communication graph.

B. Random Walk With Self-Loop Over D2D-MSN

Having defined the walk over $\mathcal{G}_m(t)$ as above, a subsequent question is, "How should the next node be selected by the

walk?" Because such walk is performed only over the communication graph, let us define $\mathcal{V}_m(t) \subseteq \mathcal{V}$ as the set of nodes that has at least one edge in $\mathcal{E}_m(t)$. Recall that $P(u,v)$ is the probability that $v \in \mathcal{V}_m(t)$ is selected as the next node when the walker is currently located at $u \in \mathcal{V}_m(t)$. Then, the walk over the communication graph forms a Markov chain with a $|\mathcal{V}_m(t)| \times |\mathcal{V}_m(t)|$ transition probability matrix $\mathbf{P}$, with $P(u,v)$ as its (u,v)-th element. Because $\mathbf{P}$ governs the movement of the walker by determining which node to visit next, we are interested in assigning its elements, $P(u,v)$.

In assigning the elements of $\mathbf{P}$, it is desirable that the corresponding Markov chain converges to its stationary distribution $\boldsymbol{\pi}$, a $1 \times |\mathcal{V}_m(t)|$ vector that satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$. The main reason behind this is because once convergence occurs, then the probability $P_{sel}(u)$ that a node u is selected by the walk is approximately equal to its stationary probability π_u, the u-th element of $\boldsymbol{\pi}$. As will be discussed in the next subsection, $P_{sel}(u)$, which is useful for eliminating the bias in the joint degree distribution's estimation, can then be calculated using π_u whose closed-form solution is readily available, as long as $\mathbf{P}$ converges to its stationary distribution, $\boldsymbol{\pi}$. To guarantee such convergence, it is known that the transition matrix $\mathbf{P}$ should be *reversible, irreducible,* and *aperiodic* [14]. We will assume that these properties hold for now and verify that the walk proposed here satisfies them later on. By such assumption, the Markov chain corresponding to $\mathbf{P}$ will always converge to its stationary distribution, *albeit* after a possibly very large number of iterations (i.e., with *slow* convergence). We then ask ourselves, "Is it possible to build a Markov chain with fast convergence to its stationary distribution?" To answer this, let us borrow the notion of *mixing time*, defined as [14]

$$\tau_{mixing}(\epsilon) := \inf(k : \max_{u \in \mathcal{V}_s(t)} |(P(u,\cdot))^k - \pi_{(\cdot)}| \leq \epsilon). \quad (2)$$

The mixing time represents the number of iterations required until the Markov chain $\mathbf{P}$ converges to its stationary distribution $\boldsymbol{\pi}$, given that a small gap $\epsilon > 0$ is allowed. Then, achieving a Markov chain with fast convergence to its stationary distribution is equal to minimizing the mixing time. Let λ_i be the i-th largest eigenvalue of $\mathbf{P}$. The mixing time of a Markov chain with transition matrix $\mathbf{P}$ is bounded by [14, Thms. 12.3-12.4]

$$\left(\frac{1}{\gamma} - 1\right) \log\left(\frac{1}{2\epsilon}\right) \leq \tau_{mixing}(\epsilon) \leq \frac{1}{\gamma} \log\left(\frac{1}{\epsilon \pi_{min}}\right), \quad (3)$$

where $\gamma := \inf_{i \geq 2} 1 - |\lambda_i|$ and $\pi_{min} := \min_u \pi_u$. From the upper and lower bounds in (3), the mixing time scales as γ^{-1}, indicating that it grows inversely-proportional to γ such that the minimal mixing time can be achieved by maximizing γ. The problem of maximizing γ subject to a symmetric $\mathbf{P}$ has been studied in [15], but is not applicable to our case since the complete knowledge regarding $\mathcal{G}_m(t)$, which characterizes all of the possible next nodes selected by the walk and is needed to assign the whole elements of $\mathbf{P}$ at once, is not readily available. In order to maximize γ for the D2D-MSN considered in this paper, we propose the following scheme, instead.

Notice that maximizing γ is equivalent to minimizing the second largest eigenvalue modulus (SLEM) $\lambda^* = \max_{i \geq 2} |\lambda_i|$. For any *stochastic* matrix $\mathbf{P}$, in which each of the row sums to unity, *i.e.*, $\mathbf{P1} = \mathbf{1}$, the smallest possible SLEM of $\lambda^* = 0$ is achieved when the Markov chain corresponds to a random walk over a complete graph in which self-loop is allowed. The transition matrix of such walk is given as $\mathbf{P}^* = \frac{1}{|\mathcal{V}_m(t)|}\mathbf{1}\mathbf{1}^T$. In this paper, however, the communication graph to be walked is not a complete graph. As a result, when located at node u, a walker may only choose either to move to one of the communication neighbors $\mathcal{N}_m(u)$ or to stay at u, such that

$$P(u,v) \begin{cases} \geq 0 & \text{if } v \in \mathcal{N}_m(u) \text{ or } v = u, \\ = 0 & \text{otherwise,} \end{cases} \quad (4)$$

holds. Since $P(u,v)$ is a conditional probability given a current node u, $\sum_v P(u,v) = 1$ such that matrix $\mathbf{P}$ is stochastic. Another imposing constraint is that the complete knowledge regarding $\mathcal{E}_m(t)$ is not known *a priori* such that when a walker is located at node u, it can only assign the transition probabilities to its neighboring communication nodes by utilizing local information at u. Notice that the u-th row of $\mathbf{P}$ represents the set of possible next nodes and their selection probabilities. Motivated by the limitation, we propose to maximize γ by minimizing $|\mathbf{P} - \mathbf{P}^*|$ through row-by-row assignment of $\mathbf{P}$ such that the constraints (4) and $\mathbf{P1} = \mathbf{1}$ are satisfied. This allows $\mathbf{P}$ to be as similar as possible to $\mathbf{P}^*$, thus minimizing its SLEM and its mixing time. The problem of assigning the u-th row of $\mathbf{P}$ when the walker is at node u then becomes

$$\arg\min_{P(u,v)} \sum_v (P(u,v) - P^*(u,v))^2 \text{ s.t. } \sum_v P(u,v) = 1. \quad (5)$$

Because the solution of (5) applies only for one currently visited node u, the walker solves (5) every time it visits a new node. Let $L(f,g) := f(u,v) - \alpha g(u,v)$ be the Lagrangian of (5), where $f(u,v) := \sum_v (P(u,v) - P^*(u,v))^2$ and $g(u,v) := \sum_v P(u,v) - 1$, while α denotes a Lagrange multiplier. By substituting $g(u,v) = 0$ into $\nabla L(f,g) = 0$, the closed-form solution of (5) can be obtained as

$$P(u,v) = \begin{cases} \frac{1}{d_m(u)+1} & \text{if } v \in \mathcal{N}_m(u) \cup \{u\}, \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

which we refer as *random walk with self-loop* (RWSL). Because all nodes in $\mathcal{V}_m(t)$ are connected by undirected edges, the communication graph is connected such that RWSL is both *irreducible* and *aperiodic* [14]. Further, by solving $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$ with the elements of $\mathbf{P}$ given in (6), we have

$$\pi_u = (d_m(u) + 1)/(2|\mathcal{E}_m(t)| + |\mathcal{V}_m(t)|). \quad (7)$$

Using (7), the *reversibility* condition $\pi_u P(u,v) = \pi_v P(v,u)$ [14] also holds for all $u, v \in \mathcal{V}_m(t)$. Since $\mathbf{P}$ is reversible, irreducible, and aperiodic, it will always converge to its stationary distribution [14], justifying our assumption.

C. Unbiased Random Walk With Self-Loop over D2D-MSN

After outlining the RWSL scheme, we ask, "How to estimate the joint degree distribution from the visited nodes?" Let $\mathcal{V}_{ij} := \{u \in \mathcal{V} : |\mathcal{N}_s(u)| = i, |\mathcal{N}_m(u)| = j\}$ be the set of nodes in $\mathcal{G}(t)$ that have i and j neighbors in the social and communication graphs, respectively. Let $1_{\mathcal{A}}(u) \in \{0,1\}$ indicates that u belongs to set $\mathcal{A}$. Then, the joint degree distribution of $\mathcal{V}$ in the original graph and $\mathcal{V}'$ visited by the walk are calculated through their sample means as

$$p_{ij} = \frac{a_{ij}}{b_{ij}} = \frac{|\mathcal{V}|^{-1} \sum_{u \in \mathcal{V}} 1_{\mathcal{V}_{ij}}(u)}{|\mathcal{V}|^{-1} \sum_{u \in \mathcal{V}} 1}, \quad (8)$$

$$p'_{ij} = \frac{a'_{ij}}{b'_{ij}} = \frac{|\mathcal{V}|^{-1} \sum_{u \in \mathcal{V}} 1_{\mathcal{V}_{ij} \cap \mathcal{V}'}(u)}{|\mathcal{V}|^{-1} \sum_{u \in \mathcal{V}} 1_{\mathcal{V}'}(u)}. \quad (9)$$

By assuming that the Markov chain corresponding to $\mathbf{P}$ already converges to its stationary distribution $\boldsymbol{\pi}$, the probability that u is selected by RWSL is given as $P_{sel}(u) := \mathbb{E}(1_{\mathcal{V}'}(u)) \approx \pi_u$ such that $\mathbb{E}(a'_{ij}) = |\mathcal{V}|^{-1} \sum_{u\in\mathcal{V}} \pi_u 1_{\mathcal{V}_{ij}}(u)$ and $\mathbb{E}(b'_{ij}) = |\mathcal{V}|^{-1} \sum_{u\in\mathcal{V}} \pi_u$ can be calculated from (9), where π_u is defined in (7). Further, using first-order approximation, $\mathbb{E}(p'_{ij}) \approx \mathbb{E}(a'_{ij})/\mathbb{E}(b'_{ij})$ and $\frac{\mathbb{E}(a'_{ij})}{\mathbb{E}(b'_{ij})} \neq \frac{a_{ij}}{b_{ij}} = p_{ij}$ can be obtained, such that we can conclude that the degree distribution p'_{ij} directly obtained from the sample mean of $\mathcal{V}'$ is *biased* with respect to the target distribution, p_{ij}, of the D2D-MSN graph $\mathcal{G}(t)$.

Since p'_{ij} estimated from the nodes collected by the walker is biased, can such bias be corrected? Notice that $\mathbb{E}(a'_{ij})$ and $\mathbb{E}(b'_{ij})$ differs from a_{ij} and b_{ij}, respectively, by a factor π_u within their respective summations. Thus, the bias can be corrected if the degree distribution p_{ij} is estimated using

$$\hat{p}_{ij} = \frac{\hat{a}_{ij}}{\hat{b}_{ij}} = \frac{\sum_{u\in\mathcal{V}'_{ij}} \frac{1}{\pi_u}}{\sum_{u\in\mathcal{V}'} \frac{1}{\pi_u}}, \tag{10}$$

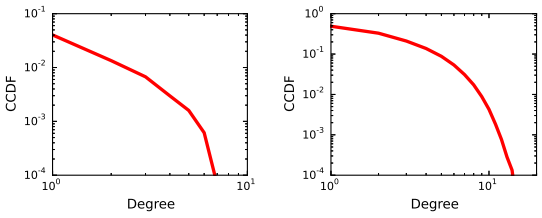
where $\mathcal{V}'_{ij} := \mathcal{V}_{ij} \cap \mathcal{V}'$ is the set of nodes in $\mathcal{V}'$ that have i and j edges in $\mathcal{E}_s(t)$ and $\mathcal{E}_m(t)$, respectively. We can verify that $\mathbb{E}(\hat{a}_{ij}) = |\mathcal{V}|^{-1} \sum_{u\in\mathcal{V}} 1_{\mathcal{V}_{ij}}(u)$ and $\mathbb{E}(\hat{b}_{ij}) = |\mathcal{V}|^{-1} \sum_{u\in\mathcal{V}} 1$, showing that $\hat{p}_{ij}$ is an *unbiased estimate* of the target distribution p_{ij}. We refer the unbiased estimator in (10) that calculates $\hat{p}_{ij}$ using the set of nodes $\mathcal{V}'$ visited by `RWSL` as an unbiased RWSL (`Unb-RWSL`) scheme.

Remark 1: Unlike the unbiased estimator for single graphs in [8], (10) provides an unbiased estimate of the *joint distribution* that fully characterizes the degree correlation in the coexisting social and communication graphs. We also note that $\mathcal{G}_s(t)$ can also be viewed as an "appending" information to $\mathcal{G}_m(t)$. Thus, the `Unb-RWSL` scheme can be *generalized* to other cases in which the social degree distribution is replaced by the distribution of device version, user gender, and so on.

IV. Numerical Results

In this section, the proposed scheme's performance is evaluated through numerical simulations in `Python`. First, we determine the structure of the social and communication graphs in realistic D2D-MSN by employing traces collected from 76 mobile phones during Sigcomm 2009 conference [16]. Note that similar results will hold for more general cases of D2D-MSns, but are not included here due to limited space. In this paper, time granularity is set to 60 seconds and we assume that a social edge occurs between two nodes if they exchange at least one message during the last hour. The resulting degree distributions in Figs. 3(a) and 3(b) indicate that (i) both social and communication degrees exhibit *exponential tail*, and (ii) the former has a smaller average degree than the latter.

Next, the performance of `Unb-RWSL` is evaluated in Fig. 2. Motivated by the results of Fig. 3, two Erdös-Rényi (ER) graph, which results in a Binomial degree distribution with exponential tail, with $n = 2 \times 10^4$ nodes and edge occurrence probabilities of $p_s^{ER} = 0.001$ and $p_m^{ER} = 0.002$ are employed for the social and communication graphs, respectively. The communication and social degrees of every node are set arbitrarily to produce an *uncorrelated* joint degree distribution, denoted as a `random` case. We employ 2×10^4 iterations with the results from the initial 10^4 iterations discarded to ensure



(a) Social graph. (b) Communication graph.

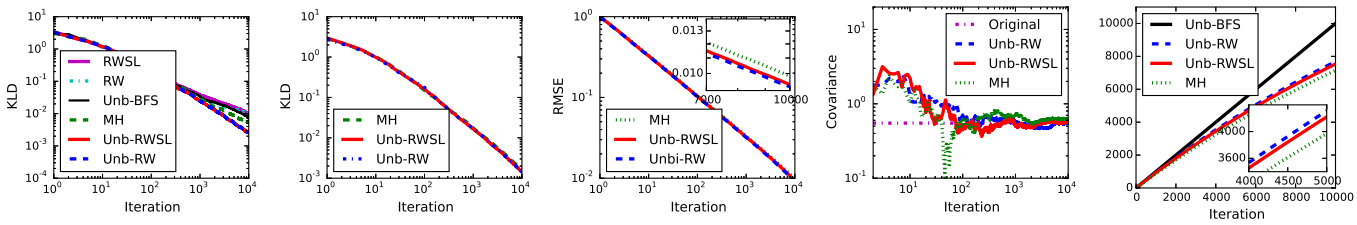
Fig. 3. Complementary CDF of the D2D-MSN graph in [16].

the walk already reaches equilibrium. The performance of the proposed scheme is compared to three existing algorithms: pure random walk (`RW`), Metropolis Hastings (`MH`) [8], and unbiased BFS (`Unb-BFS`) [7] schemes. The Kullback-Leibler divergence (KLD) [8] of the marginal communication degree distribution $p_j := \sum_i p_{ij}$ compared to its estimate $\hat{p}_j$ in Fig. 2(a) indicates that `Unb-RWSL` performs similar to pure RW employing (10), denoted as `Unb-RW`, and outperforms `MH` with a gap that increases with respect to the number of iterations. Further, a comparison between `RWSL` to `Unb-RWSL` shows that the estimator in (10) is able to correct the bias introduced by RWSL. Fig. 2(a) also shows that `Unb-BFS` achieve worse KLD performance than the `MH` and `Unb-RWSL` schemes, since it is not optimized to achieve fast convergence.

On the other hand, unlike the communication degree's KLD in Fig. 2(a), the social degree's KLD for all the examined schemes are shown to be *equal* in Fig. 2(b). To understand this, note that all the walks are performed over the communication graph. Because in the `random` case the social and communication graphs are uncorrelated, a walk over the communication graph can be viewed as a random node selection in the social graph. Consequently, `Unb-RWSL` and `Unb-RW` will both have uniform stationary distribution and similar walk structures as `MH`, such that all the three exhibits the same social degree KLD performance. Note that `Unb-BFS` [7] can only estimate the marginal degree distribution, such that its joint distribution-based statistics, e.g. KLD, root mean-squared error (RMSE), and covariance, cannot be compared to the proposed scheme.

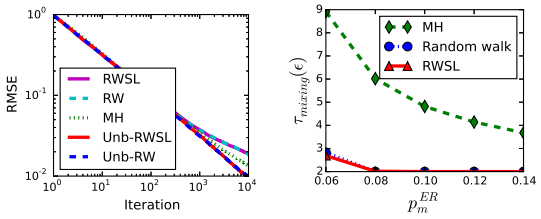
To evaluate the joint degree distribution estimation performance, the RMSE between the estimated $\hat{p}_{ij}$ to the target p_{ij} is plotted in Fig. 2(c). Similar observation as in Fig. 2(a) holds: `Unb-RWSL` performs similar to `Unb-RW` and outperforms `MH` in terms of RMSE value. This result suggests that the proposed `Unb-RWSL` is able to *well-capture* the coupling between $\mathcal{G}_s(t)$ and $\mathcal{G}_m(t)$. To further evaluate the ability of the `Unb-RWSL` in capturing the relationship between $\mathcal{G}_s(t)$ and $\mathcal{G}_m(t)$ in the joint degree distribution, let us examine Fig. 2(d), in which the covariance of the `Unb-RWSL` converges to that of $\mathcal{G}(t)$ faster than both `MH` and `Unb-RW`. The covariance fully describes the level of correlation between the social and communication degrees, verifying that the `Unb-RWSL` well captures the relationship between both graphs.

Since both `Unb-RWSL` and `Unb-RW` schemes achieve almost the same KLD and RMSE performance, then what is the advantage of employing the former? The number of distinct visited nodes *versus* iterations in Fig. 2(e) shows that the `Unb-RWSL` visits less number of distinct nodes than `Unb-RW` and `Unb-BFS`. Because the walker needs to log all the visited nodes as well as their social and communication degrees, then



(a) KLD, communication degree. (b) KLD, social degree. (c) Root mean-squared error. (d) Covariance. (e) Number of visited nodes.

Fig. 2. Joint degree distribution estimation performance of the proposed unbiased RWSL for `random` case.



(a) RMSE for `high-high` case. (b) Mixing time.

Fig. 4. RMSE and mixing time performance of the proposed scheme.

the proposed `Unb-RWSL` requires *smaller* amount of memory than the `Unb-RW` scheme, given the same number of iterations.

Next, the impact of correlation between the social and communication degrees is evaluated in Fig. 4. Here, $\mathcal{G}_s(t)$ and $\mathcal{G}_m(t)$ are set to be *fully-correlated*, *i.e.*, the node with the i-th largest communication degree also has the i-th highest social degree for all $i \leq n$, denoted as `high-high` case. Although not shown here, the KLDs of both the communication and social degrees for the `high-high` case are similar to that of the KLD of the communication degree in the `random` case in Fig. 2(a), because both graphs have identical structures and are highly-correlated. Thus, a walk over the communication graph will produce similar dynamics, indicated by the same KLD performance, over the social graph. Consequently, as depicted in Fig. 4(a), the gap between the RMSE performances of the `Unb-RWSL` and `MH` schemes, that are jointly affected by the estimation accuracy of both the social and communication degrees, will be increased, compared to that of the `random` case in Fig. 2(c). This indicates that `Unb-RWSL` has *larger* RMSE gap to `MH` scheme when the degree correlation between the social and communication graphs is high.

Finally, the mixing time of the proposed `RWSL` scheme for ER graph with $n = 200$ nodes over 500 network realizations is evaluated in Fig. 4(b). As the edge occurrence probability p_m^{ER} increases, the average communication degree will also increase, such that the transition matrix $\mathbf{P}$ of the `RWSL` scheme approaches $\mathbf{P}^*$ of the complete graph and the mixing time will be decreased. The figure also indicates that `RWSL` scheme achieves the *lowest* mixing time, similar to `RW`. This is unsurprising, since `RWSL` is similar to pure `RW`, except the former allows self-loop, *i.e.*, the walk may choose to stay at the current node. On the other hand, `MH` produces an unbiased walk with uniform stationary distribution that, unfortunately, comes with the trade-off of much larger mixing time.

V. Conclusion

We considered the problem of modeling and estimating the structure of D2D-MSNs. We defined a D2D-MSN graph that captures the social and physical relationships between users, and further employed a joint degree distribution p_{ij} to collectively characterize the structure of both graphs. An unbiased random walk with self-loop scheme on the D2D counterpart of the D2D-MSN, which achieves fast convergence to its stationary distribution as well as unbiased estimation of p_{ij}, is proposed. Numerical results showed that the proposed scheme achieves higher p_{ij} estimation accuracy than existing schemes. The estimated p_{ij} can then be used to generate a D2D-MSN graph that is applicable to the evaluation of information spreading performance, which can help determine the success of D2D-MSNs' deployment. In the future, we plan to incorporate both the MSN's and the D2D's topologies to improve the performance of the proposed scheme.

References

[1] "Social networking reaches nearly one in four around the world." http://www.emarketer.com/. Accessed: 2015-03-30.

[2] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6727–6740, 2014.

[3] S. A. Pambudi. and W. Wang, "Boundary matters: Impact of finite boundary to packet delay performance in mobile data networks," in *2014 IEEE ICC*, pp. 2748–2753, June 2014.

[4] X. Wang *et al.*, "TOSS: Traffic offloading by social network service-based opportunistic sharing in mobile social networks," in *Proc. 2014 IEEE INFOCOM*, pp. 2346–2354, IEEE, 2014.

[5] "Cisco VNI Service Adoption Forecast, 2013-2018." http://www.cisco.com/. Accessed: 2015-03-31.

[6] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, no. 1, p. 5, 2007.

[7] M. Kurant, A. Markopoulou, and P. Thiran, "Towards unbiased BFS sampling," *IEEE JSAC*, vol. 29, no. 9, pp. 1799–1809, 2011.

[8] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A case study of unbiased sampling of OSNs," in *Proc. 2010 IEEE INFOCOM*, pp. 1–9, IEEE, 2010.

[9] R. Van Der Hofstad, "Random graphs and complex networks," *Available on http://www.win.tue.nl/rhofstad/NotesRGCN.pdf*, 2009.

[10] E. Volz, "SIR dynamics in random networks with heterogeneous connectivity," *J. Math. Biol.*, vol. 56, no. 3, pp. 293–310, 2008.

[11] M. H. Hansen, W. N. Hurwitz, and W. G. Madow, "Sample survey methods and theory," 1953.

[12] H. Liu *et al.*, "Modeling/predicting the evolution trend of OSN-based applications," in *Proc. 22nd WWW*, pp. 771–780, 2013.

[13] X. Huang, J. Gao, S. V. Buldyrev, S. Havlin, and H. E. Stanley, "Robustness of interdependent networks under targeted attack," *Physical Review E*, vol. 83, no. 6, p. 065101, 2011.

[14] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*. American Math. Soc., 2009.

[15] S. Boyd, P. Diaconis, and L. Xiao, "Fastest mixing Markov chain on a graph," *SIAM review*, vol. 46, no. 4, pp. 667–689, 2004.

[16] A.-K. Pietilainen, "CRAWDAD data set thlab/sigcomm2009 (v. 2012-07-15)." Downloaded from http://crawdad.org/, July 2012.