Modeling Data Dissemination in Online Social Networks: A Geographical Perspective on Bounding Network Traffic Load

Cheng Wang
Department of Computer Science and Technology
Tongji University
Shanghai, China
chengwang@tongji.edu.cn

Shaojie Tang
Department of Information Systems
University of Texas at Dallas
Richardson, Texas
tangshaojie@gmail.com

Lei Yang
Institute of Trustworthy Networks and Systems
Tsinghua University
Beijing, China
young@tagsys.org

Yi Guo
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Kowloon, Hong Kong
guoyithomas@gmail.com

Fan Li
School of Computer Science
Beijing Institute of Technology
Beijing, China
fli@bit.edu.cn

Changjun Jiang
Department of Computer Science and Technology
Tongji University
Shanghai, China
cjjiang@tongji.edu.cn

ABSTRACT

In this paper, we model the data dissemination in online social networks (OSNs) and study the scaling laws of traffic load. We propose a three-layered system model to formulate data dissemination sessions for social applications in OSNs. The layered model consists of the physical network layer, social relationship layer, and application session layer. By analyzing mutual relevances among these three layers, we investigate the geographical distribution feature of dissemination sessions in OSNs. Based on this, we derive the traffic load of OSNs under a realistic assumption that every source sustains a data generating rate of constant order. To the best of our knowledge, this is the first work to address the issue of traffic load scaling for OSNs by modeling the social data dissemination from a layered perspective.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Network communications; Network topology; H.3.4 [**Systems and Software**]: Information networks

General Terms

Performance; Theory

Keywords

Online social networks, data dissemination, traffic load, scaling laws

MobiHoc'14, August 11–14, 2014, Philadelphia, PA, USA.

http://dx.doi.org/10.1145/2632951.2632973.

1. INTRODUCTION

Recent years have witnessed a dramatic growth of user population of online social networks (OSNs). For example, according to the report in March 2013, Facebook has 1.11 billion people using the site each month, which represents a 23 percent growth from a year earlier [3]. OSNs are organized around users [28]. Users publish their profile and any content, build links to other users in whom they are interested. OSN sites act as the basis for maintaining social relationships [18, 28]. In OSNs, the communication pattern based on users' profiles and social relationships is arguably the most important feature [18, 27]. Here, user profiles are usually personal web pages where users can post content, e.g., texts, pictures, music, and videos. For the postings of a user, some of users who follow (are interested in) this user post comments and other content as feedbacks. Thereby, data traffic emerges among OSNs. As OSNs expand rapidly, the traffic load imposed on the underlying communication networks, e.g., the Internet or different types of wireless networks, grows heavier and heavier. Then, how does such traffic load scale as the size of an OSN increases over time? This is the problem we aim to solve in this paper.

For modeling the data traffic pattern and addressing the bounds on network traffic load in OSNs, there are two key factors to consider. The first factor is the architecture of OSNs. In OSNs with the centralized architecture, users' profiles and content are uploaded into servers, and some users request from servers for the information of users of interest [28, 34]. While, in OSNs with the decentralized architecture, users run P2P clients (peers) on their hosts to browse the profiles of friends and post content. Peers form an overlay network for the purpose of collectively sharing and replicating content, serving it on behalf of offline users when needed [27, 36]. Whichever architecture is adopted, a traffic session in the OSN can be essentially modeled as a data dissemination from a source to some specific destinations. The second factor is the architecture of underlying communication network, e.g., wireless, wired, or hybrid networking. This has a significant impact on implementing routing for specific data dissemination sessions. In this paper, we are concerned with the fundamental limits on the traffic load imposed on

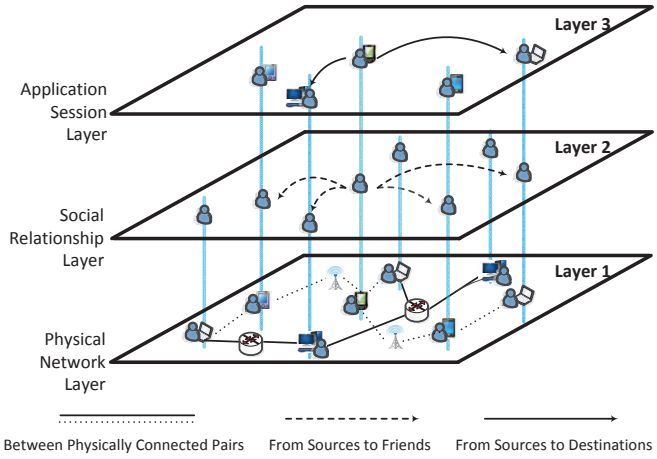

Figure 1: Layered System Model.

the underlying communication networks of OSNs, and mainly pursue the tight lower bounds that can theoretically be achieved over the communication network with the optimal architecture. We address this issue by analyzing the geographic characteristics of data dissemination sessions in OSNs, i.e., the spacial distribution of traffic sessions (the location distribution of sources and destinations).

We propose a three-layered model consisting of the physical network layer (Layer 1), social relationship layer (Layer 2), and application session layer (Layer 3), as illustrated in Fig.1. Then, for the purpose of deriving the spacial distribution of traffic sessions depending on users' geographical distribution, we adopt two steps to clarify the correlation between Layer 3 and Layer 1: Firstly, we start with dredging the correlation between Layer 2 and Layer 1, i.e., the relevance between the formation scheme of users' social relationships and the distribution model of users' geographical locations. We propose an innovative and practical *population-based social formation model* to build the links from Layer 1 to Layer 2, and validate its practicality based on a dataset of the online social network called Brightkite. Secondly, we analyze the correlation between Layer 3 and Layer 2, i.e., the relevance between the traffic pattern for a specific social application and the topology of users' social relationship network, assuming users constantly intend to deliver information to other users with whom they associate (friends or followers). By these two steps, we compute the bound on the aggregate travel distance over which a data packet in a session is transmitted from the source to its destinations. Furthermore, we derive the traffic load of the social network under a realistic assumption that all sources sustain the data generating rates of the constant order.

To emulate the inhomogeneous geographical distribution of users in real-life OSNs, we introduce the shotnoise Cox process [5, 29] to model the deployment of users on Layer 1. Under such a model, we present the general density function of social relationship distribution and the general form of traffic load bounds in OSNs, which can serve as the basis for obtaining further an in-depth understanding on scaling laws of traffic load in OSNs (Theorem 1). In particular, we derive the explicit traffic load scaling for the OSNs with a special Layer 1, where users are homogeneously distributed (Theorem 2). The result shows that the traffic load for social-broadcast is monotonically nonincreasing in the range $\left[\Theta(n), \Theta(n^2)\right]$ for the clustering exponents of both friendship degree and friendship formation (Section 5.3.3), where n is the number of users in the OSN. It can highlight the differences from results for data dissemination in conventional communication networks. Furthermore, when a specific underlying communication network is introduced, our results on travel distances can also play an important role in analyzing some system performances, e.g., network capacity and latency. To the best of our knowledge, this is the first work to address scaling laws for the traffic load of OSNs by modeling the data dissemination in social networks.

The rest of this paper is organized as follows. We provide the motivation and problem formulation of this work in Section 2, and propose the system model in Section 3. In Section 4, we give the main results on general system model, and derive the representative results for a special model where users are distributed homogeneously on the physical network layer in Section 5. Finally, we draw a conclusion, and make a discussion about the future work in Section 6.

2. PRACTICAL MOTIVATION AND PROBLEM FORMULATION

2.1 Related Work and Our Motivation

Online social networks (OSNs) allow hundreds of millions of the Internet users worldwide to produce and consume content, and provide users with the access to the unprecedented large-scale information repository [15]. They play an important role in the information diffusion by increasing the spread of novel information and diverse viewpoints, and have shown their power in many situations [6,26]. In the research community of OSNs, there are some representative topics that have been extensively studied, such as detecting popular topics [25], modeling information diffusion [30], and identifying influential spreaders [8], and so on. Most existing work mainly focused on the information diffusion scheme over overlay relationship networks of users in social network sites/services (SNSs).

Meanwhile, as SNSs become increasingly popular for information exchange, the traffic generated by social applications rapidly expands [2]. As an argument, according to a report of Shareaholic [4], between November 2012 and November 2013, social media referral traffic from the top five social media sites increased by 111% while search traffic from the top five search engines had decreased by 6%. Thereby, besides the analysis of information diffusion schemes in overlay social networks like in the literature [6, 8, 25, 26, 30, 37], an in-depth understanding of the impact of increasing traffic generated by OSNs on underlying communication networks, e.g., the Internet, is also necessary for evaluating current OSN systems, optimizing network architectures and the deployment of servers for OSNs, and even designing future OSNs. One of main challenges in addressing this issue is to propose practical modeling and effective analyzing methods for data dissemination of OSNs implemented in underlying communication networks, since OSNs change both information propagation schemes and traffic session patterns in communication networks due to the involvement of overlay social relationships and users' preferences and decisions. Accordingly, in this paper, we aim at modeling data dissemination in OSNs, and analyzing the traffic load for OSNs imposed on the underlying communication networks.

2.2 Formulation of Traffic Load for OSNs

In a communication network, the sum of products of bits and the distances over which they are carried, i.e., the so-called *traffic load*, is a valuable indicator of the network's transport capacity [16, 19]. In this paper, we mainly study the traffic load scaling of data dissemination in OSNs.

Considering an OSN, denoted by $\mathbb{N}$, consisting of n users, denote the set of all users by $\mathcal{U} = \{u_i\}_{i=1}^{n}$; let a subset $\mathcal{S} = \{u_{\mathcal{S},k}\}_{k=1}^{n_s} \subseteq$

$\mathcal{U}$ denote the set of all sources, where $|\mathcal{S}| = n_s$. Before giving the formal definition for the traffic load of the social network $\mathbb{N}$ in Section 2.2.3, we introduce two key conceptions: *data arrival process for users* and *transport distance of messages*.

2.2.1 Data Arrival Process for Users

Individual Data Arrival Process: The temporal behavior of messages arriving at a user in an OSN has been addressed by analyzing some real-life OSNs, [7, 32]. For example, Perera et al. [32] developed a software architecture that uses a Twitter application program interface (API) to collect the tweets sent to specific users. They indicated that the arrival process of new tweets to a user can be modeled as a Poisson Process. In this work, we just make it as an empirical argument for assuming the data arrival for a user as a data source to follow a Poisson Process. We define a rate vector

$$\mathbf{\Lambda}_{\mathcal{S}} = (\lambda_{\mathcal{S},1}, \lambda_{\mathcal{S},2}, \cdots, \lambda_{\mathcal{S},n_s}),$$

where $\lambda_{\mathcal{S},k}$ is the rate of the Poisson Process at user $u_{\mathcal{S},k}$ (for $k = 1, 2, \cdots, n_s$), and is called *data arrival/generating rate*. Here, we make a reasonable and practical assumption that $\lambda_{\mathcal{S},k} = \Theta(1)$ for $k = 1, 2, \cdots, n_s$.

Data Arrival Rate Distribution: The results of interest in this paper fall within the scope of scaling laws issue. Therefore, although in practise, the individual data arrival/generating process may depend on many factors such as the user's location, it is appropriate to note at this point that the specific distribution of data arrival rate has no impact on the results (in order sense) as long as it holds that $\lambda_{\mathcal{S},k} = \Theta(1)$ for $k = 1, 2, \cdots, n_s$. This is the reason why we do not make an intensive study of the specific distribution of data arrival rates.

2.2.2 Transport Distance of Messages

Denote a data dissemination session from a source $u_{\mathcal{S},k}$ by an ordered pair $\mathbb{D}_{\mathcal{S},k} =< u_{\mathcal{S},k}, \mathcal{D}_{\mathcal{S},k} >$, where $\mathcal{D}_{\mathcal{S},k}$ is the set of all destinations of $u_{\mathcal{S},k}$. Define a distance vector

$$\mathbf{D}_{\mathcal{S}} = (d_{\mathcal{S},1}, d_{\mathcal{S},2}, \cdots, d_{\mathcal{S},n_s}),$$

where $d_{\mathcal{S},k}$ represents the *transport distance* over which the message for session $\mathbb{D}_{\mathcal{S},k}$ is successfully transported from the source $u_{\mathcal{S},k}$ to the destinations.

The transport distance depends on the specific architectures of underlying communication networks and routing strategies. While, in this work, we aim to present a general result which is independent of the specific network communication architecture due to the diversity of real-life communication networks. Then, we focus on investigating the achievable lower bound on the transport distance. The lower bound achieved by the optimal communication strategy can serve as a reasonable metric to quantify and bound the transport distances over different underlying communication networks. For a session $\mathbb{D}_{\mathcal{S},k}$, when the underlying communication network is involved, the problem to obtain the optimal transport distance can be reduced to the Euclidean Steiner Tree Problem spanning the set $\{u_{\mathcal{S},k}\} \cup \mathcal{D}_{\mathcal{S},k}$: Given $|\{u_{\mathcal{S},k}\} \cup \mathcal{D}_{\mathcal{S},k}|$ nodes in the plane, the goal is to connect them by lines of minimum total length in such a way that any two nodes may be interconnected by line segments either directly or via other nodes (relay devices in the communication network) and line segments [13].

2.2.3 Traffic Load for Online Social Networks

In the OSN $\mathbb{N}$, define the traffic load for a dissemination session, say $\mathbb{D}_{\mathcal{S},k}$, as $\widetilde{\mathbf{L}}_{\mathbb{N},\mathcal{S}}(\mathbb{D}_{\mathcal{S},k}) = \lambda_{\mathcal{S},k} \cdot d_{\mathcal{S},k}$. Furthermore, the total traffic load for data dissemination from all sources in $\mathcal{S}$ can be explicitly defined as

$$\mathbf{L}_{\mathbb{N},\mathcal{S}} = \mathbf{\Lambda}_{\mathcal{S}} * \mathbf{D}_{\mathcal{S}}, \tag{1}$$

where $*$ is an inner product.

Next, we give a basic lemma to derive the traffic load by each session.

LEMMA 1. *Under the assumption that $\lambda_{\mathcal{S},k} = \Theta(1)$ for $k = 1, 2, \cdots, n_s$, it holds that*

$$\widetilde{\mathbf{L}}_{\mathbb{N},\mathcal{S}}(\mathbb{D}_{\mathcal{S},k}) = \Omega\left(|\operatorname{EMST}\left(\{u_{\mathcal{S},k}\} \cup \mathcal{D}_{\mathcal{S},k}\right)|\right),$$

$$\mathbf{L}_{\mathbb{N},\mathcal{S}} = \Omega\left(\sum\nolimits_{k=1}^{n_s} |\operatorname{EMST}\left(\{u_{\mathcal{S},k}\} \cup \mathcal{D}_{\mathcal{S},k}\right)|\right),$$

where $\operatorname{EMST}(\cdot)$ *denotes the Euclidean minimum spanning tree over a set.*

PROOF. A minimum spanning tree is a feasible but not usually optimal solution to the Steiner tree problem. The Steiner ratio is the largest possible ratio between the total length of a minimum spanning tree and the total length of a minimum Steiner tree [11]. In the Euclidean Steiner tree problem, the Steiner ratio is conjectured to be $\frac{2}{\sqrt{3}}$. The conjecture is still open [17], although earlier claims of a proof were stated in [11]. Whatever the specific value of the Steiner ratio is, it is definitely a constant. Therefore, when we only care the order of final results in this paper, the value of the Steiner ratio has no impact in order sense. As a result, we can use the total length of the Euclidean spanning tree of session $\mathbb{D}_{\mathcal{S},k}$, i.e., the Euclidean spanning tree over the set $\{u_{\mathcal{S},k}\} \cup \mathcal{D}_{\mathcal{S},k}$, to measure the order of the total length of the optimal transport distance for session $\mathbb{D}_{\mathcal{S},k}$.

Combining with the fact that $\lambda_{\mathcal{S},k} = \Theta(1)$ for all k, we can prove this lemma. □

We remark that the lower bounds in Lemma 1 can be achieved by the optimal underlying communication network architecture and communication strategy.

3. LAYERED SOCIAL NETWORK MODEL

For each session, the geographical distribution of the source and destination(s) plays a key role in generating the traffic load. Then, it is critical to analyze the correlation between the spatial distribution of sessions and geographical distribution of users.

To address this issue, we introduce a three-layer model, consisting of the physical network layer (Layer 1), social relationship layer (Layer 2), and application session layer (Layer 3), as illustrated in Fig.1, for modeling data dissemination in OSNs. The basic procedure of modeling the correlations between the geographical distributions of sessions and users can be described as follows:

1. At first, based on the users' geographical distribution in the physical network layer, we build the relationships among users to form the social relationship layer.

2. Then, based on the formed social relationship layer, we model the geographical distribution of traffic sessions.

3.1 Layer 1: Physical Network Deployment

To embody the uneven population distribution in real-life OSNs, we introduce the cluster random model.

We consider the random network consisting of a random number N (with $\mathbf{E}(N) = n$)[1] users distributed over a square region of area

[1]Throughout the paper, we let $\mathbf{E}[X]$ denote the mean and variance of a random variable X, respectively.

$S := S(n)$. To avoid border effects, we consider wraparound conditions at the network edges. That is, the network area is assumed to be the surface of a two-dimensional Torus, denoted by $\mathcal{O}$. The network physical extension $\sqrt{S}$ is allowed to scale with the average number of nodes, since this is expected to occur in many growing systems.

The so-called clustering random model (CRM) can be denoted by $\mathbb{N}(n, S; \mathbb{C}(m), g(\cdot))$, where the expected number of centers, say m, and the clustering function, say $g(\cdot)$, will be defined later.

To emulate the clustering behavior of users' distribution in real-life OSNs, it is necessary to introduce the shotnoise Cox process (SNCP, [5, 29]).

3.1.1 Shotnoise Cox Process

An SNCP can be described by the following construction: First, specify a point process $\mathbb{C}(m)$ of cluster centers, whose positions are denoted by $\mathcal{C} = \{c_j\}_{j=1}^{M}$, where M is a random number with average $\mathbf{E}(M) = m$, and the center points c_j are also called parent points. Then, each center point c_j in turn generates a point process of nodes whose intensity at position X is given by $\rho_j \cdot \kappa(c_j, X)$, where $\rho_j \in (0, \infty)$ and $\kappa(c_j, \cdot)$, called kernel or shot, is a dispersion density function. The nodes generated by each center are referred to as offspring points. The overall node process $\mathbb{P}$ is then given by the superposition of the individual processes generated by the cluster centers. The conditional local intensity at X of the resulting SNCP is

$$\mathbf{d}(X) = \sum\nolimits_j \rho_j \cdot \kappa(c_j, X),$$

where $\mathbf{d}(X)$ is a random field conditionally over all (ρ_j, c_j). Then, the node process $\mathbb{P}$ forms an inhomogeneous Poisson point process with intensity function $\mathbf{d}(\cdot)$. Let $\mathcal{X} = \{X_i\}_{i=1}^{N}$ denote the collection of nodes positions in a given realization of the SNCP.

Under the above assumptions on the kernel shape, the quantity ρ_j equals the average number of nodes generated by cluster center c_j. We assume that all cluster centers generate on average the same number of nodes, hence it holds that $\rho_j = \rho = \frac{n}{m}$. In this work, we let ρ scale with n as well. That is, clusters are expected to grow in size as the number of nodes increases.

In this paper, we restrict ourselves to kernels $\kappa(c_j, \cdot)$ that are invariant under both translation and rotation. The kernel $\kappa(c_j, X) = \kappa(|X - c_j|)$ depends only on the Euclidean distance $|X - c_j|$ of point X from the cluster center c_j. Moreover, we assume that $\kappa(c_j, \cdot)$ is a summable, non-increasing, bounded and continuous function, and $\int_{\mathcal{O}} \kappa(c_j, X) dX = 1$. Following a common normalizing method, the kernels can be specified by first defining a non-increasing, bounded and continuous *clustering function* $g(s)$ and then normalizing it over the network area $\mathcal{O}$:

$$\kappa(c_j, X) = \frac{g(|X - c_j|)}{\int_{\mathcal{O}} g(|Y - c_j|) dY} \tag{2}$$

3.1.2 Clustering Random Model $\mathbb{N}(n, S; \mathbb{C}(m), g(\cdot))$

We define the *clustering random model* (CRM) by specifying the point process $\mathbb{C}(m)$ of cluster centers.

- When $\mathbb{C}(m)$ is a homogeneous Poisson process (HPP) of intensity $\lambda_c = m/S$ over $\mathcal{O}$, we call such a CRM *Poisson clustering random model*.
- When m are deployed in a regular grid pattern, we call such a CRM *Regular clustering random model*.

For both models, we define $l_c = \sqrt{S/m}$. This quantity represents the length of edges of the square where the expected number of contained cluster centers is equal to 1. We call the case that $l_c = O(1)$ *cluster-dense regime*, and call the case $l_c = \omega(1)$ *cluster-sparse regime*, in which l_c tends to infinity as n increases.

To simplify the description, we assume that the number of nodes is exactly n, and denote the set of nodes by $\mathcal{V} = \{v_k\}_{k=1}^{n}$, which has no impact on our results in order sense.

3.2 Layer 2: Social Relationship Formation

We propose a density-aware social relationship formation model, called *population-based social formation model*. We will clarify the advantages of this model later in Section 3.2.2.

Let $\mathcal{D}(u, r)$ denote the disk centered at a point u with radius r in the deployment region $\mathcal{O}$, and let $N(u, r)$ denote the number of nodes contained in $\mathcal{D}(u, r)$.

3.2.1 Population-based Social Formation Model

For a node $v_k \in \mathcal{V}$, construct its friendship set of q_k, $q_k \geq 1$, nodes/friends, say $\mathcal{F}_k$, by the following procedure:

1. Zipf's Degree Distribution of Social Relations: Assume that the number of friends (or followers) of a particular node $v_k \in \mathcal{V}$, denoted by q_k, follows a Zipf's distribution [24, 33], i.e.,

$$\Pr(q_k = l) = \left(\sum\nolimits_{j=1}^{n-1} j^{-\gamma}\right)^{-1} \cdot l^{-\gamma}. \tag{3}$$

From Eq.(3), we can observe that the degree distribution above depends on the specific network size (the number of users n). We will give a numerical validation based on Brightkite dataset for the Zipf's degree distribution in Appendix B.2.

We notice that the correlations between the users' degree distribution and graphical distribution should not be ignored, although we simplify it in this work. We will address this issue in the future work.

2. Population-Based Formation of Social Relations: Making the position of node v_k as the *reference point*, choose q_k *points* independently on the torus region $\mathcal{O}$ according to a probability distribution with density function:

$$f_{v_k}(X) = \Phi_k(S, \beta) \cdot (\mathbf{E}[N(v_k, |X - v_k|)] + 1)^{-\beta}, \tag{4}$$

where the random variable $X := (x, y)$ denotes the position of a selected point in the deployment region, $|X - v_k|$ denotes the Euclidean distance between point X and node v_k, $\beta \in [0, \infty)$ represents the clustering exponent of friendship formation; the coefficient $\Phi_k(S, \beta) > 0$ depends on β and S (the area of deployment region), satisfying that:

$$\Phi_k(S, \beta) \cdot \int_{\mathcal{O}} (\mathbf{E}[N(v_k, |X - v_k|)] + 1)^{-\beta} \, dX = 1. \tag{5}$$

3. Nearest-Principle Position of Friends/Followers: Let $\mathcal{A}_k = \{p_{k_i}\}_{i=1}^{q_k}$ denote the set of these q_k points. Let v_{k_i} be the nearest node to p_{k_i}, for $1 \leq i \leq q_k$ (ties are broken randomly). Denote the set of these q_k nodes by $\mathcal{F}_k = \{v_{k_i}\}_{i=1}^{q_k}$. We call point p_{k_i} the *anchor point* of v_{k_i}, and define a set $\mathcal{P}_k := \{v_k\} \cup \mathcal{A}_k$.

Throughout this paper, we use $\mathbb{P}(\delta, \gamma, \beta)$ to denote the population-based social model.

3.2.2 Advantages of Population-Based Social Model

After Kleinberg [20] proposed a distance-based social model relating geographical distance and social friendship, Liben-Nowell et al. [23] introduced the rank-based model, where the probability of befriending a particular person is inversely proportional to the power of the number of closer people. They validated the practicality of rank-based model by analyzing the data of an online social network, the LiveJournal online community. They pointed out that the

Table 1: Notations for Exponents

Notation	Definition
$\delta \in [0, \infty)$	clustering exponent of node distribution
$\gamma \in [0, \infty)$	clustering exponent of friendship degree
$\beta \in [0, \infty)$	clustering exponent of friendship formation

weakness of distance-based models lies in that for a particular user, it underestimates the friendship probability of the distant nodes in the low-density region, when the geographical distribution of users is inhomogeneous in common occurrence.

The rank-based model states that the friendship probability depends on both the geographic distance and node density. Following this observation, by modifying the rank-based model, we propose the distance & density-aware population-based social model. The population-based model is more convenient and systematic for addressing the bounds on aggregate travel distances. Anchor points, defined in Step 3 of Section 3.2.1, are effectively introduced, in order to ensure the independence of length of certain Euclidean spanning trees, thus makes it convenient to bound the total length, e.g., the proof of Lemma 2. However, under the rank-based model where the friendships are directly built over nodes without anchor points, the corresponding independence cannot be guaranteed, which usually brings the difficulty on the theoretical rigor.

We will provide a numerical validation based on Brightkite dataset for the population-based social formation model in Appendix B.3.

3.3 Layer 3: Application Session Construction

After the social layer is formed, social sessions can be defined according to the specific applications: For the *social-unicast*/*social-multicast*, the source node delivers message to one/multiple selected friend(s).

For the *social-broadcast*, the source node broadcasts message to all its friends, such as tweets in Twitter and posts in Facebook. Accordingly, we can define other session patterns based on the definitions of corresponding non-social sessions, such as *social-anycast* [22] and *social-manycast* [9].

4. MAIN RESULTS FOR GENERAL CLUSTERING RANDOM MODEL

In this paper, we focus on the extended networks [12, 14]) with a constant average node density. Specifically, for the physical network layer under the clustering random model (CRM), denoted by $\mathbb{N}(n, S; \mathbb{C}(m), g(\cdot))$, we let $S = n$ to study the extended model $\mathbb{N}(n, n; \mathbb{C}(m), g(\cdot))$. Furthermore, for the clustering function $g(\cdot)$, we define it as

$$g(s) := \min\left\{1, s^{-\delta}\right\}$$

for the extended model $\mathbb{N}(n, n; \mathbb{C}(m), g(\cdot))$, where $\delta \in [0, \infty)$ is the clustering exponent of node distribution. Note that when $\delta = 0$, the model degenerates into the homogeneous random extended network [12, 14].

To facilitate the reader, we have reported in Table 1 a collection of frequently-used system parameters.

4.1 General Density Function

In the clustering random model $\mathbb{N}(n, n; \mathbb{C}(m), g(\cdot))$, we construct a set of $q + 1$ *points*, denoted by $\mathcal{P} = \{X_i\}_{i=0}^{q}$, by the following procedure:

Step 1. Select arbitrarily a point from $\mathcal{O}$ as the first one in $\mathcal{P}$, denoted by X_0.

Step 2. Making point X_0 as the *reference point* denoted by O', select independently other q points at random according to the probability distribution with density function as described in Eq.(4) (Let $v_k := X_0$).

LEMMA 2. *Making the point X_0 as the reference point O', the distribution of points in $\mathcal{A} := \{X_i\}_{i=1}^{q}$ follows the probability distribution with the density function*

$$f_{X_0}(X) = \frac{\left[\int_{\mathcal{D}(X_0, |X - X_0|)} \mathbf{d}(Y)\, dY + 1\right]^{-\beta}}{\int_{\mathcal{O}} \left[\int_{\mathcal{D}(X_0, |Z - X_0|)} \mathbf{d}(Y)\, dY + 1\right]^{-\beta} dZ} \tag{6}$$

where

$$\mathbf{d}(Y) = \frac{n}{m} \cdot \sum\nolimits_{c_j \in \mathcal{C}} \frac{\min\left\{1, |Y - c_j|^{-\delta}\right\}}{\int_{\mathcal{O}} \min\left\{1, |Z - c_j|^{-\delta}\right\} dZ}.$$

4.2 Traffic Load Distribution

4.2.1 Main Results on Traffic Load

We can use the set $\mathcal{P} = \{X_i\}_{i=0}^{q}$ to represent a social traffic session, where X_0 acts as the source node and $\mathcal{A} := \{X_i\}_{i=1}^{q}$ is the set of all destinations of X_0. Then, we have

THEOREM 1. *For the traffic load by the session with the set $\mathcal{P}$, denoted by $\widetilde{\mathbf{L}}$, it holds that: when $q = \omega(1)$, with high probability $1 - o(1/\hat{N})$, it follows that*

$$\widetilde{\mathbf{L}} = \Omega(\Upsilon), \tag{7}$$

where

$$\Upsilon : \left[\sqrt{q} \cdot \int_{\mathcal{O}} \sqrt{f_{X_0}(X)} dX, \sqrt{q} \cdot \int_{\mathcal{O}} \sqrt{f_{X_0}(X)} dX + \bar{L}\right]^{2}, \tag{8}$$

$f_{X_0}(X)$ is defined in Eq.(6) of Lemma 2, and

$$\bar{L} = \min\left\{L \,\middle|\, \int_{\mathcal{D}(X_0, L)} f_{X_0}(X) dX = \Omega\left(\min\left\{\frac{\log \hat{N}}{q}, 1\right\}\right)\right\} \tag{9}$$

with a given parameter $\hat{N} : (1, n]$ denoting the number of nodes with degree of order $\omega(1)$.

4.2.2 Proof of Theorem 1

Let $\mathrm{EMST}(\mathcal{P})$ and $\mathrm{EMST}(\mathcal{A})$ denote the Euclidean minimum spanning trees of $\mathcal{P} = \{X_i\}_{i=0}^{q}$ and $\mathcal{A} := \{X_i\}_{i=1}^{q}$, respectively.

We first give a lemmas providing the total edge length of Euclidean minimum spanning trees of $\mathcal{P} = \{X_i\}_{i=0}^{q}$, denoted by $|\,\mathrm{EMST}(\mathcal{P})|$.

LEMMA 3. *When $q = \omega(1)$, it holds that:*
▷ With probability 1,

$$|\,\mathrm{EMST}(\mathcal{A})| = \Theta\left(\sqrt{q} \cdot \int_{\mathcal{O}} \sqrt{f_{X_0}(X)} dX\right), \tag{10}$$

where $f_{X_0}(X)$ is defined in Eq.(6) of Lemma 2.

[2]We use the term $f(n) : [\underline{\phi}(n), \overline{\phi}(n)]$ to represent $f(n) = \Omega(\underline{\phi}(n))$ and $f(n) = O(\overline{\phi}(n))$; and use $f(n) : (\underline{\phi}(n), \overline{\phi}(n))$ to represent $f(n) = \omega(\underline{\phi}(n))$ and $f(n) = o(\overline{\phi}(n))$.

$\triangleright$ *With high probability* $1 - o(1/\hat{N})$,

$$|\operatorname{EMST}(\mathcal{P})| : \left[|\operatorname{EMST}(\mathcal{A})|, |\operatorname{EMST}(\mathcal{A})| + \bar{L}\right], \quad (11)$$

where $\bar{L}$ *is defined in Eq.(9).*

PROOF. First of all, the cost of an edge (X_i, X_j) is given by

$$\Psi(|X_i - X_j|) = |X_i - X_j|.$$

That is, the exponent σ in Lemma 8 equals 1. In addition, $\Psi(x)$ is a monotonically increasing function.

Let L denote the distance between the center O and reference point. Then, under the center-clustering random model $\mathbb{N}^1(n, S; g(\cdot))$, by Eq.(4) and Eq.(5), the density function is specified into Eq.(13). Then, by Lemma 8, we get that

$$|\operatorname{EMST}(\mathcal{A})| = \Theta\left(\sqrt{q} \cdot \textstyle\int_{\mathcal{O}} \sqrt{f_{X_0}(X)} dX\right).$$

It is straightforward that

$$|\operatorname{EMST}(\mathcal{P})| = \Omega\left(|\operatorname{EMST}(\mathcal{A})|\right).$$

On the other hand, let $\underline{L}$ denote the smallest distance from the points in $\mathcal{A}$ to point X_0. Then,

$$\left(1 - \textstyle\int_{\mathcal{D}(X_0, \underline{L})} f_{X_0}(X) dX\right)^q = o(1).$$

That is,

$$\int_{\mathcal{D}(X_0, \underline{L})} f_{X_0}(X) dX = \omega\left(\frac{1}{q}\right).$$

Thus, $\underline{L} \leq \bar{L}$, where $\bar{L}$ is defined in Eq.(9), which completes the proof. Note that we deliberately relax the upper bound of $\underline{L}$ as in Eq.(9) in order to ensure Eq.(11) to hold with uniformly high probability for $\Theta(n)$ Euclidean spanning trees [31]. □

Finally, combining with Lemma 1, we complete the proof.

4.3 Special Results for Single-Center Model

According to the fact that there usually exist a finite number of dense areas in real-life network applications, we mainly focus on the finite-centers clustering random model (FC-CRM), denoted by $\mathbb{N}(n, n; \mathbb{C}(c), g(\cdot))$, where $c > 0$ is a constant.

Since we aim to derive scaling laws of aggregate travel distances, we only care the order of the parameters and performance metrics. Then, it holds that the finite-centers clustering random model is indeed similar to a clustering random model with exactly one center, which is called single-center clustering random model (SC-CRM) here, in terms of scaling behavior. Denote such a SC-CRM by $\mathbb{N}(n, n; \mathbb{C}(1), g(\cdot))$.

We construct a $\mathbb{N}(n, n; \mathbb{C}(1), g(\cdot))$ by the following procedure: First, making a center of $\mathcal{O}$ as the center point, denoted by O. Then, the center point O generates a point process of nodes whose local intensity at position X is given by

$$\mathbf{d}(X) = n \cdot \kappa(O, X) = n \cdot \frac{\min\left\{1, |Y - O|^{-\delta}\right\}}{\int_{\mathcal{O}} \min\left\{1, |Z - O|^{-\delta}\right\} dZ}. \quad (12)$$

where $\kappa(O, \cdot)$ is the dispersion density function. Then, we get the probability density function as follows.

COROLLARY 1. *Making the point* X_0 *as the reference point* O'*, the distribution of points in* $\mathcal{A} := \{X_i\}_{i=1}^q$ *follows the probability with the density function*

$$f_{X_0}(X) = \frac{\left[\int_{\mathcal{D}(X_0, |X - X_0|)} \mathbf{d}(Y)\, dY + 1\right]^{-\beta}}{\int_{\mathcal{O}} \left[\int_{\mathcal{D}(X_0, |Z - X_0|)} \mathbf{d}(Y)\, dY + 1\right]^{-\beta} dZ} \quad (13)$$

where

$$\mathbf{d}(Y) = n \cdot \frac{\min\{1, |Y - O|^{-\delta}\}}{\int_{\mathcal{O}} \min\{1, |Z - O|^{-\delta}\} dZ}.$$

Combining Corollary 1 and Theorem 1, we can obtain the traffic load by a session under the single-center clustering random model.

5. A CASE STUDY: TRAFFIC LOAD FOR OSNS WITH HOMOGENEOUS DISTRIBUTION OF USERS

In this section, in order to obtain an explicit result, we specifically reduce the complexity from three dimensions $(\delta, \gamma, \beta) \in [0, \infty)^3$ to two dimensions $(\gamma, \beta) \in [0, \infty)^2$ by letting $\delta = 0$. In this case of extremely weak clustering behavior, the physical layer degenerates into the homogeneous random network model [12, 16], where $\mathbf{d}(Y) \equiv \Theta(1)$.

From Eq.(3), we get the degree distribution as follows:

$$\Pr(q_k = l) = \begin{cases} \Theta(l^{-\gamma}), & \gamma > 1; \\ \Theta(\frac{1}{\log n} \cdot l^{-1}), & \gamma = 1; \\ \Theta(n^{\gamma - 1} \cdot l^{-\gamma}), & 0 \leq \gamma < 1. \end{cases} \quad (14)$$

We can reasonably assume that all nodes behave as source nodes, i.e., $\mathcal{S} = \mathcal{V}$. Then, the description of social broadcast session $\mathbb{D}_{\mathcal{S},k}$ can be further simplified. That is, the session initiated from node v_k is denoted by $\mathbb{D}_k$.

5.1 Distribution of Anchor Points

For each dissemination session $\mathbb{D}_k$ initiated by the source v_k, we can get the distribution of anchor points directly using Corollary 1.

LEMMA 4. *When the clustering exponent* $\delta = 0$*, for a session* $\mathbb{D}_k$ *under the population-based social model* $\mathbb{P}(\delta = 0, \gamma, \beta)$*, the anchor points of the friends of source* v_k *follows the distribution of density function:*

$$f_{v_k}(X) = \begin{cases} \Theta\left(\left(|X - v_k|^2 + 1\right)^{-\beta}\right), & \beta > 1; \\ \Theta\left(\left(\log n \cdot \left(|X - v_k|^2 + 1\right)\right)^{-1}\right), & \beta = 1; \\ \Theta\left(n^{\beta - 1} \cdot \left(|X - v_k|^2 + 1\right)^{-\beta}\right), & 0 \leq \beta < 1. \end{cases}$$

By using Lemma 4, we can get the following result.

LEMMA 5. *For a social-broadcast session* $\mathbb{D}_k$ *under the model* $\mathbb{P}(\delta = 0, \gamma, \beta)$*, it holds that:*

$$\mathbf{E}[|X - v_k|] = \begin{cases} \Theta(1), & \beta > 3/2; \\ \Theta(\log n), & \beta = 3/2; \\ \Theta(n^{\frac{3}{2} - \beta}), & 1 < \beta < 3/2; \\ \Theta(\sqrt{n}/\log n), & \beta = 1; \\ \Theta(\sqrt{n}), & 0 \leq \beta < 1. \end{cases} \quad (15)$$

5.2 Social-Broadcast Sessions

Under the population-based social model, we denote a social-broadcast session $\mathbb{D}_k$ by the set $\{v_k\} \cup \mathcal{F}_k$, where v_k is the source and each element in $\mathcal{F}_k = \{v_{k_i}\}_{i=1}^{q_k}$, say v_{k_i}, is the nearest *node* to the corresponding *anchor point* p_{k_i} in $\mathcal{A}_k = \{p_{k_i}\}_{i=1}^{q_k}$. Recall that $\mathcal{P}_k = \{v_k\} \cup \mathcal{A}_k$, we can get the following Lemma 6 for spanning trees over $\mathbb{D}_k$.

LEMMA 6. *For a social-broadcast session* $\mathbb{D}_k$ *with* $q_k = \omega(1)$ *under the social formation model* $\mathbb{P}(\delta = 0, \gamma, \beta)$*, with probability* 1*, it holds that*

$$|\operatorname{EMST}(\mathcal{A}_k)| = \Theta(L_{\mathcal{P}}(\beta, q_k)),$$

Table 2: $H(\gamma,\beta)$ Depending on Exponents γ and β

γ	$H(\gamma,\beta)$
$\gamma>2$	$\begin{cases}\Theta(n), & \beta>2;\\ \Theta(n\cdot\log n), & \beta=2;\\ \Theta(n^{2-\frac{\beta}{2}}), & 1<\beta<2;\\ \Theta(n^{3/2}/\sqrt{\log n}), & \beta=1;\\ \Theta(n^{3/2}), & 0\le\beta<1.\end{cases}$
$\gamma=2$	$\begin{cases}\Theta(n\cdot\log n), & \beta\ge2;\\ \Theta(n^{2-\frac{\beta}{2}}), & 1<\beta<2;\\ \Theta(n^{3/2}/\sqrt{\log n}), & \beta=1;\\ \Theta(n^{3/2}), & 0\le\beta<1.\end{cases}$
$3/2<\gamma<2$	$\begin{cases}\Theta(n^{3-\gamma}), & \beta\ge2\gamma-2;\\ \Theta(n^{2-\frac{\beta}{2}}), & 1<\beta<2\gamma-2;\\ \Theta(n^{3/2}/\sqrt{\log n}), & \beta=1;\\ \Theta(n^{3/2}), & 0\le\beta<1.\end{cases}$
$\gamma=3/2$	$\begin{cases}\Theta(n^{3/2}), & \beta>1;\\ \Theta(n^{3/2}\cdot\sqrt{\log n}), & \beta=1;\\ \Theta(n^{3/2}\cdot\log n), & 0\le\beta<1.\end{cases}$
$1<\gamma<3/2$	$\Theta(n^{3-\gamma})$
$\gamma=1$	$\Theta(n^2/\log n)$
$0\le\gamma<1$	$\Theta(n^2)$

and then

$$|\operatorname{EMST}(\mathcal{P}_k)|=\Omega(L_{\mathcal{P}}(\beta,q_k)),$$

where

$$L_{\mathcal{P}}(\beta,q_k)=\begin{cases}\Theta\left(\sqrt{q_k}\right), & \beta>2;\\ \Theta\left(\sqrt{q_k}\cdot\log n\right), & \beta=2;\\ \Theta\left(\sqrt{q_k}\cdot n^{1-\frac{\beta}{2}}\right), & 1<\beta<2;\\ \Theta\left(\sqrt{q_k}\cdot\sqrt{\frac{n}{\log n}}\right), & \beta=1;\\ \Theta\left(\sqrt{q_k}\cdot\sqrt{n}\right), & 0\le\beta<1.\end{cases}\tag{16}$$

PROOF. From Theorem 3, it follows that with probability 1,

$$|\operatorname{EMST}(\mathcal{A}_k)|=\Theta(L_{\mathcal{P}}(\beta,q_k))\text{ for }q_k=\omega(1),$$

where $L_{\mathcal{P}}(\beta,q_k)$ is defined in Eq.(16). Combining with the fact that $|\operatorname{EMST}(\mathcal{P}_k)|\ge|\operatorname{EMST}(\mathcal{A}_k)|$, we get the lemma. □

5.3 Main Result on Traffic Load

5.3.1 Lower Bound on Traffic Load

The main result in this section can be described by the following theorem.

THEOREM 2. *Under the social model $\mathbb{P}(\delta=0,\gamma,\beta)$, the traffic load for data dissemination in OSN $\mathbb{N}$, denoted by $\mathbf{L}_{\mathbb{N}}$, is of*

$$\mathbf{L}_{\mathbb{N}}=\Omega\left(H(\gamma,\beta)\right),$$

where $H(\gamma,\beta)$ is defined in Table.2.

5.3.2 Proof of Theorem 2

The bound depends on the bound on $\sum_{k=1}^{n}|\operatorname{EMST}(\mathcal{P}_k)|$. We firstly give a basic lemma for the final proof.

LEMMA 7. *For all social-broadcast sessions $\mathbb{D}_k$ ($k=1,2,\cdots,n$) under the social formation model $\mathbb{P}(\delta=0,\gamma,\beta)$, with high probability, the lower bounds on $\sum_{k=1}^{n}|\operatorname{EMST}(\mathcal{P}_k)|$ hold as described in Table.3.*

PROOF. Let N_l denote the number of sessions with l destinations. First of all, to simplify the proof, we let

$$N_l=n\cdot\Pr(q_k=l)=n\cdot\left(\sum_{j=1}^{n-1}j^{-\gamma}\right)^{-1}\cdot l^{-\gamma},$$

which has no impact on the analysis in order sense according to laws of larger numbers. Based on all $\mathbb{D}_{\mathcal{S},k}$, define two sets

$$\mathcal{K}^1:=\{k|q_k=\Theta(1)\}\text{ and }\mathcal{K}^\infty:=\{k|q_k=\omega(1)\}.$$

Then,

$$\sum\nolimits_{k=1}^{n}|\operatorname{EMST}(\mathcal{P}_k)|=\underline{\Sigma}^1+\underline{\Sigma}^\infty,\tag{17}$$

where

$$\underline{\Sigma}^1=\sum_{k\in\mathcal{K}^1}|\operatorname{EMST}(\mathcal{P}_k)|,\ \underline{\Sigma}^\infty=\sum_{k\in\mathcal{K}^\infty}|\operatorname{EMST}(\mathcal{P}_k)|.$$

First, we consider $\underline{\Sigma}^1$. Since for $q_k=\Theta(1)$, it holds that

$$|\operatorname{EMST}(\mathcal{P}_k)|=\Theta(|X-v_k|),$$

then we have

$$\underline{\Sigma}^1=\sum\nolimits_{k\in\mathcal{K}^1}|X-v_k|.$$

For $k\in\mathcal{K}^1$, define a sequence of random variables $\xi_k^1:=|X-v_k|/\sqrt{n}$ having finite mean:

$$\mathbf{E}[\xi_k^1]=\mathbf{E}[|X-v_k|]/\sqrt{n},$$

where $\mathbf{E}[|X-v_k|]$ is presented in Lemma 5. Then,

$$\underline{\Sigma}^1=\Theta\left(\sqrt{n}\cdot\sum\nolimits_{k\in\mathcal{K}^1}\xi_k^1\right).$$

Thus, by Lemma 9, with probability 1,

$$\sum\nolimits_{k\in\mathcal{K}^1}\xi_k^1=\Theta\left(|\mathcal{K}^1|\cdot\mathbf{E}[|X-v_k|/\sqrt{n}]\right),$$

where $|\mathcal{K}^1|$ denotes the cardinality of $\mathcal{K}^1$. Thus, we get that with probability 1,

$$\underline{\Sigma}^1=\Theta\left(|\mathcal{K}^1|\cdot\mathbf{E}[|X-v_k|]\right).\tag{18}$$

Next, we consider $\underline{\Sigma}^\infty$. For $k\in\mathcal{K}^\infty$, all random variables $|\operatorname{EMST}(\mathcal{P}_k)|$ are *independent*; moreover, from Lemma 6, with probability 1,

$$|\operatorname{EMST}(\mathcal{P}_k)|=\Omega(L_{\mathcal{P}}(\beta,q_k)),$$

where $L_{\mathcal{P}}(\beta,q_k)$ is defined in Eq.(16). Thus, with probability 1,

$$\underline{\Sigma}^\infty\ge\sum\nolimits_{l:(1,n]}n\cdot\left(\sum\nolimits_{j=1}^{n-1}j^{-\gamma}\right)^{-1}\cdot l^{-\gamma}\cdot L_{\mathcal{P}}(\beta,l).\tag{19}$$

Finally, combining Eqs.(17), (18) and (19), we complete the proof. □

Then, we begin to prove Lemma 2. Since

$$\sum\nolimits_{k=1}^{n}q_k=\Theta\left(\sum\nolimits_{l=1}^{n-1}n\cdot\Pr(q_k=l)\cdot l\right),$$

Table 3: Lower Bounds on $\sum_{k=1}^{n} |\operatorname{EMST}(\mathcal{P}_k)|$

β \ γ	$\gamma > 3/2$	$\gamma = 3/2$	$1 < \gamma < 3/2$	$\gamma = 1$	$0 \le \gamma < 1$
$\beta > 2$	$\Omega(n)$	$\Omega(n \log n)$	$\Omega(n^{\frac{5}{2}-\gamma})$	$\Omega(n^{3/2}/\log n)$	$\Omega(n^{3/2})$
$\beta = 2$	$\Omega(n \cdot \log n)$	$\Omega(n \cdot (\log n)^2)$	$\Omega(\log n \cdot n^{\frac{5}{2}-\gamma})$	$\Omega(n^{3/2})$	$\Omega(n^{3/2} \cdot \log n)$
$1 < \beta < 2$	$\Omega(n^{2-\frac{\beta}{2}})$	$\Omega(n^{2-\frac{\beta}{2}} \cdot \log n)$	$\Omega(n^{\frac{7}{2}-\gamma-\frac{\beta}{2}})$	$\Omega(n^{(5-\beta)/2}/\log n)$	$\Omega(n^{(5-\beta)/2})$
$\beta = 1$	$\Omega(n^{3/2}/\sqrt{\log n})$	$\Omega(n^{3/2} \cdot \sqrt{\log n})$	$\Omega(n^{3-\gamma}/\sqrt{\log n})$	$\Omega(n^2/(\log n)^{3/2})$	$\Omega(n^2/\sqrt{\log n})$
$0 \le \beta < 1$	$\Omega(n^{3/2})$	$\Omega(n^{3/2} \cdot \log n)$	$\Omega(n^{3-\gamma})$	$\Omega(n^2/\log n)$	$\Omega(n^2)$

we get that $\sum_{k=1}^{n} q_k = Q(\gamma)$, where

$$Q(\gamma) = \begin{cases} \Theta(n), & \gamma > 2; \\ \Theta(n \log n), & \gamma = 2; \\ \Theta\left(n^{3-\gamma}\right), & 1 < \gamma < 2; \\ \Theta\left(n^2/\log n\right), & \gamma = 1; \\ \Theta\left(n^2\right), & 0 \le \gamma < 1. \end{cases} \tag{20}$$

Moreover, for all $v_k \in \mathcal{V}$,

$$\mathbf{E}\left[|v_{k_i} - p_{k_i}|\right] = \Theta\left(\int_0^{\sqrt{n}} x \cdot e^{-\pi \cdot x^2} dx\right),$$

that is,

$$\mathbf{E}\left[|v_{k_i} - p_{k_i}|\right] = \Theta(1).$$

Thus, according to Lemma 9, with high probability, it follows that

$$\sum\nolimits_{k=1}^{n} \sum\nolimits_{i=1}^{q_k} |v_{k_i} - p_{k_i}| = \Theta\left(\sum\nolimits_{k=1}^{n} q_k\right). \tag{21}$$

Combining with Lemma 7, we get that: for all social-broadcast sessions $\mathbb{D}_k$, $k = 1, 2, \cdots, n$, under the social model $\mathbb{P}(\delta = 0, \gamma, \beta)$, with high probability,

$$\sum\nolimits_{k=1}^{n} |\operatorname{EMST}(\mathbb{D}_{\mathcal{S},k})| = \Omega(H(\gamma, \beta)),$$

where $H(\gamma, \beta)$ is described in Table.2.

Finally, according to Lemma 1, we can complete the proof of Theorem 2.

5.3.3 *Explanation of Results*

At first, we discuss the impacts of clustering exponents of friendship degree and friendship formation, i.e., γ and β, on the traffic load. The traffic load for social-broadcast is monotonically nonincreasing in the range $\left[n, n^2\right]$ for both γ and β. An intuitive explanation can be made as follows: A larger clustering exponent of friendship degree γ can limit the number of friends of each user into a smaller upper bound with high probability, then leads to a lower traffic load; a larger clustering exponent of friendship formation β makes the friends more closer to each user with high probability, then possibly reduces the total transmission distance of each social-broadcast session, finally also leads to a lower traffic load.

Importantly, we notice that this work should be regarded as the first step for investigating the traffic load under the population-based model. Here, we only take into account the case that $\delta = 0$, where the population-based model degenerates to that similar to distance-based model [20,21]. The advantages of population-based model cannot be sufficiently highlighted for such a special model, indeed. It would be a significant future work to clarify the relationships between the general clustering exponent and traffic load in 3-dimensional parameter space, i.e., $(\delta, \gamma, \beta) \in [0, \infty)^3$.

6. CONCLUSION AND FUTURE WORK

We have presented a three-layered architecture to model the data dissemination in online social networks (OSNs), i.e., the physical network layer (Layer 1), social relationship layer (Layer 2), and application session layer (Layer 3). By analyzing mutual relevances among these three layers, we get the geographical distribution characteristics of dissemination sessions in OSNs. Based on this, we presented the density function of general social relationship distribution and the general form of traffic load bounds for OSNs, and derived the tight lower bounds on traffic load of data dissemination in the OSNs.

Much work still remains. Under the profile & social-based information dissemination pattern, a subsequent traffic session from a source is usually triggered by the previous session from another source. While, we have focused exclusively the data arrival model where the correlations of data generating processes at sources have been ignored. Besides this, we have not investigate the correlation between the degree distribution and the geographical distribution of the user in our model. Even for our proposed model, we only provided the explicit result for the model with homogeneous geographical distribution of users. This cannot still highlight sufficiently the characteristics of real-life OSNs and the advantages of the proposed population-based formation model.

Acknowledgements

The research of authors is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61202383, 61370192 and 60903151, the National Basic Research Program of China (973 Program) under Grant No. 2010CB328101, Shanghai Rising-Star Program under Grant No. 14QA1403700, the Program for New Century Excellent Talents in University (NCET) under Grant No. NCET-12-0414, the Natural Science Foundation of Shanghai under Grant No. 12ZR1451200, the Integrated Project for Major Research Plan of the National Natural Science Foundation of China under Grant No. 91218301, the Research Fund for the Doctoral Program of Higher Education of China (RFDP) under Grant No. 20120072120075, the Shanghai Foundation for Development of Science and Technology under Grant No. 11JC1412800, the National Key Technology R&D Program under Grant No. 2012BAH15F03, and the Beijing Natural Science Foundation under Grant No. 4122070.

7. REFERENCES

[1] Brightkite dataset. `http://snap.stanford.edu/data/loc-brightkite.html`.

[2] Cisco visual networking index: Global mobile data traffic forecast update, 2013-2018.

http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white\textunderscorepaper\textunderscorerc11-520862.html.

[3] Number of active users at facebook over the years. http://news.yahoo.com/number-active-users-facebook-over-230449748.html.

[4] Shareaholic's search traffic vs. social referrals report. https://blog.shareaholic.com/search-traffic-social-referrals-12-2013/.

[5] G. Alfano, M. Garetto, and E. Leonardi. Capacity scaling of wireless networks with inhomogeneous node density: Upper bounds. *IEEE Journal on Selected Areas in Communications*, 27(7):1147–1157, 2009.

[6] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proc. ACM WWW 2012*.

[7] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *Proc. ACM IMC 2009*.

[8] J. Borge-Holthoefer, A. Rivero, and Y. Moreno. Locating privileged spreaders on an online social network. *Physical Review E*, 85(6):066123, 2012.

[9] C. Carter, S. Yi, P. Ratanchandani, and R. Kravets. Manycast: exploring the space between anycast and multicast in ad hoc networks. In *Proc. ACM MobiCom 2003*.

[10] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: Friendship and mobility: User movement in location-based social networks. In *Proc. ACM SIGKDD 2011*.

[11] D. Du and F. Hwang. A proof of the Gilbert-Pollak conjecture on the Steiner ratio. *Algorithmica*, 7(1):121–135, 1992.

[12] M. Franceschetti, O. Dousse, D. Tse, and P. Thiran. Closing the gap in the capacity of wireless networks via percolation theory. *IEEE Transactions on Information Theory*, 53(3):1009–1018, 2007.

[13] E. Gilbert and H. Pollak. Steiner minimal trees. *SIAM Journal on Applied Mathematics*, 16(1):1–29, 1968.

[14] G. Grimmett. *Percolation*. Springer, 1999.

[15] A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: A survey. *SIGMOD Record*, 42(2):17, 2013.

[16] P. Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2):388–404, 2000.

[17] A. O. Ivanov and A. A. Tuzhilin. The steiner ratio Gilbert–Pollak conjecture is still open. *Algorithmica*, 62(1-2):630–632, 2012.

[18] L. Jin, Y. Chen, T. Wang, P. Hui, and A. V. Vasilakos. Understanding user behavior in online social networks: A survey. *IEEE Communications Magazine*, 51(9):144–150, 2013.

[19] A. Jovicic, P. Viswanath, and S. R. Kulkarni. Upper bounds to transport capacity of wireless networks. *IEEE Transactions on Information Theory*, 50(11):2555–2565, 2004.

[20] J. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 2000.

[21] B. Latané, J. Liu, A. Nowak, M. Bonevento, and L. Zheng. Distance matters: Physical space and social impact. *Personality and Social Psychology Bulletin*, 21(8):795–805, 1995.

[22] V. Lenders, M. May, and B. Plattner. Density-based anycast: a robust routing strategy for wireless ad hoc networks. *IEEE/ACM Transactions on Networking*, 16(4):852–863, 2008.

[23] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS of the United States of America*, 102(33):11623–11628, 2005.

[24] C. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[25] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proc. ACM SIGMOD 2010*.

[26] M.Cataldi, C. L. Di, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proc. MDMKDD 2010*.

[27] G. Mega, A. Montresor, and G. P. Picco. Efficient dissemination in decentralized social networks. In *Proc. IEEE P2P 2011*.

[28] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. ACM IMC 2007*.

[29] J. Møller. Shot noise cox processes. *Advances in Applied Probability*, 35(3):614–640, 2003.

[30] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proc. ACM SIGKDD 2012*.

[31] M. Penrose. A strong law for the longest edge of the minimal spanning tree. *The Annals of Probability*, 27(1):246–260, 1999.

[32] R. Perera, S. Anand, K. Subbalakshmi, and R. Chandramouli. Twitter analytics: Architecture, tools and analysis. In *Proc. IEEE MILCOM 2010*.

[33] A. Sala, H. Zheng, B. Y. Zhao, S. Gaito, and G. P. Rossi. Brief announcement: revisiting the power-law degree distribution for social graph analysis. In *Proc. ACM PODC 2010*.

[34] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger. Understanding online social network usage from a network perspective. In *Proc. ACM IMC 2009*.

[35] J. Steele. Growth rates of Euclidean minimal spanning trees with power weighted edges. *The Annals of Probability*, pages 1767–1787, 1988.

[36] J. Sun, X. Zhu, and Y. Fang. A privacy-preserving scheme for online social networks with efficient revocation. In *Proc. IEEE INFOCOM 2010*.

[37] F. Wang, H. Wang, K. Xu, J. Wu, and X. Jia. Characterizing information diffusion in online social networks with linear diffusive model. In *Proc. IEEE ICDCS 2013*.

[38] D. Williams. *Probability with martingales*. Cambridge university press, 1991.

APPENDIX

A. USEFUL LEMMAS

LEMMA 8 (MINIMAL SPANNING TREE [35]). *Let X_i, $1 \leq i < \infty$, denote independent random variables with values in $\mathbb{R}^d$, $d \geq 2$, and let M_n denote the cost of a minimal spanning tree of a complete graph with vertex set $\{X_i\}_{i=1}^n$, where the cost of an*

edge (X_i, X_j) is given by $\Psi(|X_i - X_j|)$. Here, $|X_i - X_j|$ denotes the Euclidean distance between X_i and X_j and Ψ is a monotone function. For bounded random variables and $0 < \sigma < d$, it holds that as $n \to \infty$, with probability 1, one has

$$M_n \sim c_1(\sigma, d) \cdot n^{\frac{d-\sigma}{d}} \cdot \int_{\mathbb{R}^d} f(X)^{\frac{d-\sigma}{d}} dX,$$

provided $\Psi(x) \sim x^\sigma$, where $f(X)$ is the density of the absolutely continuous part of the distribution of the $\{X_i\}$.

LEMMA 9 (KOLMOGOROV'S STRONG LLN [38]). *Let $\{X_n\}$ be an i.i.d. sequence of random variables having finite mean: for $\forall n$, $\mathbf{E}[X_n] < \infty$. Then, a strong law of large numbers (LLN) applies to the sample mean:*

$$\bar{X}_n \xrightarrow{a.s.} \mathbf{E}[X_n],$$

where $\xrightarrow{a.s.}$ denotes almost sure convergence.

B. A VALIDATION BASED ON BRIGHTKITE DATASET

In this section, we provide the validations of the adopted degree distribution model and the proposed population-based social formation model using Brightkite users' dataset [10].

B.1 Brightkite Dataset

Brightkite was created in 2007. It was once a location-based social networking service provider where users shared their locations by "checking-in" function, [1]. The friendship network was collected using their public API, and consists of $58,228$ nodes and $214,078$ edges. The Brightkite users' dataset in [10] collected a total of $4,491,143$ checkins of these users over the period from April 2008 to October 2010. It provides each user with the incoming and outgoing friend lists.

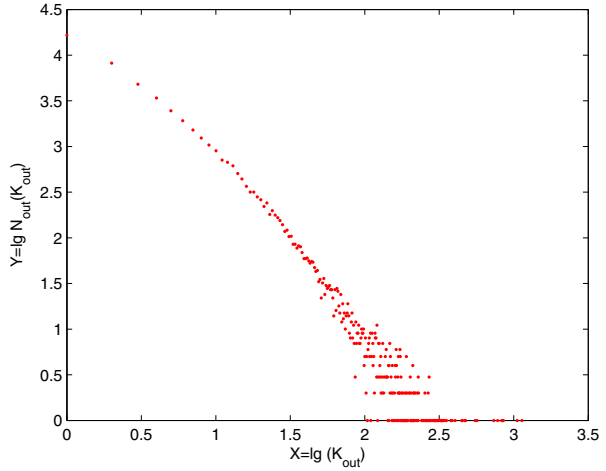

Figure 2: Social Degree Distribution of Brightkite Users.

B.2 Degree Distribution of Brightkite Users

Recall that we assume that the number of friends of a particular node $v_k \in \mathcal{V}$, denoted by q_k, follows a Zipf's distribution [24], i.e.,

$$\Pr(q_k = l) = \left(\sum_{j=1}^{n-1} j^{-\gamma}\right)^{-1} \cdot l^{-\gamma}.$$

We validate the Zipf's degree distribution of social relations by investigating the negative linear correlation between

$$Y := \lg N_{out}(K_{out}) \text{ and } X := \lg K_{out},$$

where K_{out} represents an outgoing degree, and $N_{out}(K_{out})$ denotes the number of the users with the outgoing degree K_{out}.

In the Brightkite dataset [10], the relationship between Y and X is described as Fig.2. It shows that the relationship be approximated to a line segment with negative slope, which basically matches our proposed model.

B.3 Validation of Population-Based Model

Let $d(u, v)$ denote the distance between user u and user v; let $\mathcal{D}(u, v)$ denote the disk centered at u with a radius $d(u, v)$; and let $N(u, v)$ denote the number of nodes in the disk $\mathcal{D}(u, v)$. Furthermore, we define a variable

$$\mathbf{I}(u, v) = \mathbf{1} \cdot \{v \text{ is a friend of } u\}.$$

We validate the power-law degree distribution of social formation by investigating the *negative linear correlation* between Y and X, where $X := \lg N$ with N denoting a number of nodes, and

$$Y := \lg \left(\frac{\sum_{<u,v> \in \mathcal{E}} \mathbf{1} \cdot \{N(u, v) = N\}}{\sum_{u,v \in \mathcal{V}} \mathbf{1} \cdot \{N(u, v) = N\}} \right),$$

with $\mathcal{V}$ and $\mathcal{E}$ denoting the set of all users and the set of all social links, respectively.

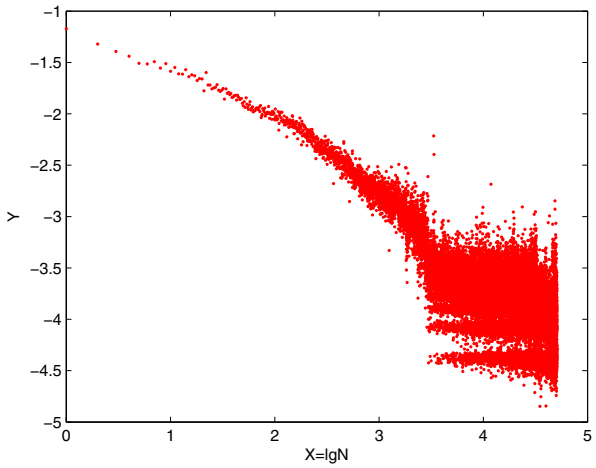

Figure 3: Population-based Social Probability Distribution of Brightkite Users.

In the Brightkite dataset [10], the relationship between Y and X is described as Fig.3. It shows that the relationship tendency is approximated very coarsely to a line segment with negative slope. The experimental result also basically validates our proposed model, although it does not perfectly match. The main reason of mismatch lies in the fact that: (1) The locations of users in the dataset are coarse-grained, and are indeed estimated in the experiments. This reduces the accuracy of experiments. (2) Based on this dataset, more than 90% results fall within the part with $X > 3$. The accumulation of experimental errors here leads to a "bloated" tail in the validation graph.