A Model of Information Diffusion in Interconnected Online Social Networks

A model of information diffusion in interconnected online social networks

ROSSANO GAETA

Online social networks (OSN) have today reached a remarkable capillary diffusion. There are numerous examples of very large platforms people use to communicate and maintain relationships. People also subscribe to several OSNs, e.g, people create accounts on Facebook, Twitter, and so on. This phenomenon leads to online social internetworking (OSI) scenarios where users who subscribe to multiple OSNs are termed as *bridges*. Unfortunately, several important features make the study of information propagation in an OSI scenario a difficult task, e.g., correlations in both the structural characteristics of OSNs and the bridge interconnections among them, heterogeneity and size of OSNs, activity factors, cross-posting propensity, etc. In this paper we propose a directed random graph-based model that is amenable to efficient numerical solution to analyze the phenomenon of information propagation in an OSI scenario; in the model development we take into account heterogeneity and correlations introduced by both topological (correlations among nodes degrees and among bridge distributions) and user-related factors (activity index, cross-posting propensity). We first validate the model predictions against simulations on snapshots of interconnected OSNs in a reference scenario. Subsequently, we exploit the model to show the impact on the information propagation of several characteristics of the reference scenario, i.e., size and complexity of the OSI scenario, degree distribution and overall number of bridges, growth and decline of OSNs in time, and time-varying cross-posting users propensity.

1. INTRODUCTION

Nowadays, online social networks (OSN) have become a key medium for information diffusion and amplification. There are countless daily examples of news, rumors, emotions that use OSNs to spread and reach a large number of people throughout the whole world.

A fraction of people subscribe to multiple OSNs (we term them as *bridges*) and represent the key elements of an online social internetworking scenario (OSI). Bridges have the capability to cross-post information received from a given OSN to a subset of the other OSNs they are part of; they actually allow the information to use additional pathways to diffuse and to reach more users.

Modeling and analysis of information spreading in an OSI scenario are crucial to understand the impact of different system parameters on the diffusion process. Nevertheless, it is a difficult task because OSNs are numerous, large, heterogeneous, and interconnected. From the topological point of view, correlations in both the structural characteristics of OSNs and the bridge interconnections among them pose a difficult challenge to both modeling and efficient analysis. Moreover, user-related factors such as users activity, and cross-posting propensity interact with this complex topological scenario thus making the analysis even harder.

Simulation and/or measurement-based analysis have been carried out but are either very complex or partial due to the size of each OSN and to the complexity arising from the bridge-based interconnection of several such complex systems.

Paper contribution

The main contribution of this paper is the development of a tractable mathematical model for the analysis of information propagation across multiple OSNs. This problem is highly complex because of many factors: structural heterogeneities of each OSN, bridges that interconnect them, and heterogeneities associated with users such as their activity patterns, interests, and propensities.

In particular, we propose a directed generalized random graph-based modeling framework to study the number of accounts in all OSNs composing the OSI scenario that receive an

information originating from an OSN. In this case, information travels inside the originating OSN and crosses OSN boundaries thanks to cross-postings operated by bridges.

To this end, we include the relevant topological feature of each OSN by representing them as directed generalized random graphs with correlations among nodes degrees; we also describe the bridge interconnection among OSNs by means of the probabilities a node belongs to any pair of OSNs that compose the OSI scenario. As far as the user-related factors are concerned, we consider the information spreading process originating from a randomly chosen node in a given OSN taking into account users activity, interest, and propensity to cross-post the information to other OSNs.

All the model parameters are OSN specific therefore heterogeneity of the OSI scenario is readily included in the model specification. The numerical model solution complexity allows to consider an OSI scenario composed of several large scale OSNs.

Validation is performed by running simulations on real snapshots of interconnected OSNs in an OSI scenario [Buccafurri et al. 2013]: we will prove that the model predictions are reliable and accurate on large scale topologies we obtain through a graph magnification operation. We successively use the model to investigate the impact of several system parameters to the size of diffusion of information in a reference OSI scenario. In particular, we consider the impact of the:

— size and complexity of the OSI scenario,
— degree distribution and overall number of bridges,
— growth and decline of OSNs in time, and
— time-varying cross-posting users propensity.

We observed interesting relationships between degree distribution and overall number of bridges on the actual information propagation as well as non-monotonic behavior when popularity of OSNs evolves in time.

The paper is organized as follows: Section 2 describes our system and formalizes all relevant concepts, Section 3 presents the mathematical derivation of the generalized random graph model we developed, Section 4 contains model validation through simulation, as well as model exploitation, Section 5 discusses related works, and in Section 6, we draw conclusions and outline ongoing activities that extend the current work.

2. SYSTEM DESCRIPTION

In this section we set up the terminology and notation we adopt to describe and formalize the topological organization of the OSI scenario we consider; we then illustrate the information spreading process and the user-related factors we include in our modeling and analysis.

2.1. OSI topological organization

The OSI scenario we consider is composed of a set of X OSNs. Users may subscribe to a subset of them; we denote a user who subscribed to multiple OSNs as a *bridge* while we denote a user who owns only one account in only one OSN as an *island*. In the following, we denote as $\mathcal{X}$ the index set $\mathcal{X} = \{1, 2, \ldots, X\}$ and we use lowercase letters x, y, and w to identify OSNs in $\mathcal{X}$.

We consider users that subscribe to an OSN and create one *account*; for the sake of simplicity, we assume a user creates only one account in a given OSN. Users establish *contacts* between their account and others in the OSN (*outgoing* contacts); contacts from other accounts are termed as *incoming* contacts. Then, $\forall x \in \mathcal{X}$ we denote as d_x a pair of nonnegative integers $d_x = (i_x, o_x)$ to represent that an account in OSN x has i_x (o_x) incoming (outgoing) contacts from (to) other accounts; we term d_x as the account *degree* and i_x (o_x) as the in-degree (out-degree).

The interconnection among OSNs is realized by means of bridges co-located in multiple OSNs. Bridges represent only a fraction of the whole set of accounts in an OSN; the inter-

connection *strength* depends on the fraction of accounts in each pair of OSNs in $\mathcal{X}$ as well as on the bridges degree in both.

2.1.1. Formalization. We describe each OSN in $\mathcal{X}$ as a directed generalized random graph [Avrachenkov et al.]; its topology is described by the degree distribution $\{p(d_x)\}$ that represents a partition of accounts in OSN x that is based on the number of contacts they have.

Unfortunately, degree distribution is not enough to account for the many existing correlations in the OSN topologies, e.g., [Mislove et al. 2007; Krishnamurthy et al. 2008; Cha et al.]. To this end, we complement the topology description of OSN x by considering probability distribution $\{p(d'_x|d_x)\}$, i.e., the probability that a randomly chosen degree d_x account has an outgoing contact pointing to a degree d'_x account.

The formalization of the interconnection among OSNs is realized by the concept of *strength* between any pair of OSNs. Formally, $\forall x, y \in \mathcal{X}$ such that $x \neq y$, we denote the strength of the connection of OSN x towards OSN y through d_x and d_y degree accounts as $b(d_x, d_y)$, that is, the joint probability that a randomly chosen account in OSN x has degree d_x and it is at the same time a degree d_y account in OSN y. Please note that:

— strength is not a symmetrical concept therefore, in general, $b(d_x, d_y) \neq b(d_y, d_x)$;
— strength of connection between OSN x and itself is clearly $b(d_x, d'_x) = p(d_x)$ if $d_x = d'_x$ and 0 otherwise.

2.2. Information propagation process

We focus on an information originating from a randomly chosen account in OSN x; we denote this triggering account as the information *origin*. We consider diffusion occurring thanks to forwarding actions taken by accounts to share information with their outgoing contacts. To this end, both the origin and an intermediate account select outgoing contacts to forward them the information. Actual selection of outgoing contacts depends on many parameters; in the following we describe what we believe are the most influential.

— **Information content**: users forward information if they are interested in the content [Liu et al. ; Wen et al. 2015]. Interest is also an OSN-specific concept, e.g., gossip news could be of little interest to Linkedin users.
— **Information age**: previous works, e.g., [Ye and Wu 2010; Cha et al. 2008] suggest that users forward information with a propensity that is a function of the information age and it is also content specific.
— **Forwarding mechanism offered by OSNs**: the forwarding mechanism is OSN dependent. For instance, in Twitter tweets and re-tweets are meant to be received by all outgoing contacts of a forwarding account while Facebook allows one to define subsets of contacts to be included in the information sharing. Also in this case, interest in information might have an impact on the size of the subset of contacts with whom to share the information.

Another important issue in the diffusion of an information is related to users activity [Rejaie et al. 2010; Torkjazi et al. 2009; Liu et al. 2013; Ribeiro 2014]. Indeed, all the forwarding actions we previously described can be taken only by *active* users, i.e., users who access their account and carry on activities on it.

Finally, in an OSI scenario information can cross the OSN boundaries and diffuse in an OSN other than the originating one thanks to bridges who cross-post the information [Reza Farahbakhsh and Crespi 2015; Ottoni et al.]. Indeed, the process of switching OSN while sharing information is becoming a basic functionality provided by many platforms and OSN aggregators. Nevertheless, some bridges might operate without the aid of these facilitating tools, e.g., tools which enables users to connect to multiple OSNs with a single authentication, therefore might be less prone to forward the information.

2.2.1. Formalization. We abstract the complex process of selection of outgoing contacts to share the information with by assuming that accounts either forward or discard the information based on the result of flipping a coin whose weight is OSN-specific, degree-dependent, and age-dependent [González et al.]. In particular, a degree d_x account in OSN $x \in \mathcal{X}$ forwards the information to a random subset of its outgoing contacts in the same OSN with probability $f_{d_x}^{(t)}$ (that we denote as the *forwarding probability*) where t is the information age, i.e., the number of forwarding steps taken starting from the origin. With probability $1 - f_{d_x}^{(t)}$ the information is discarded, i.e., it is forwarded to 0 outgoing contacts.

If accounts forward the information to their outgoing contacts then information forwarding is probabilistically carried out to a random subset of them. To this end, we model this process as accounts flipping a coin before sending or forwarding the information to an outgoing contact. Also in this case, the weight of this coin is OSN-specific, degree-dependent, and age-dependent. In particular, a degree d_x account in OSN $x \in \mathcal{X}$ forwards the information to one of its outgoing contacts in the same OSN with probability $q_{d_x}^{(t)}$ (that we denote as the *contact selection probability*). Please note that both $\{f_{d_x}^{(t)}\}$ and $\{q_{d_x}^{(t)}\}$ *are not* probability distributions.

We represent activity of users of OSN x as the fraction of active accounts over the total number of accounts. We allow this probability to be degree-dependent [Gonzalez et al.] and we denote it as α_{d_x}. Here we do not further detail if activity is defined as the daily active users or the monthly active users; the use of one activity index with respect to the other would only affect the actual values of α_{d_x} used for analysis.

To represent information cross-posting, we define *crossing propensity* β_{d_x,d_y}, i.e., the probability that a degree d_x account in OSN x takes a forwarding action for the information using his/her account in OSN y wherein the degree is d_y. Crossing propensity can also be interpreted as an indirect measure of how easily the forwarding action can be taken in an OSI scenario. Please note that β_{d_x,d_y} needs not to be symmetric, i.e., $\beta_{d_x,d_y} \neq \beta_{d_y,d_x}$, and that in the same OSN x propensity is maximum, i.e., $\forall d_x, d'_x, \beta_{d_x,d'_x} = 1$. Lastly, please note that crossing propensities do not form a probability distribution. To ease reading, all the paper notation is summarized in Table I.

3. SYSTEM MODEL

In this section we present the generalized random graph-based modeling to represent information propagation in an OSI scenario. In Section 3.1 we derive the mean number of contacts that are *potentially* reachable from the origin through the OSN networks. We then proceed with the derivation of the mean number of *actual* contacts receiving the information in Section 3.2.

3.1. Potential propagation

The network of accounts and contacts among them in OSN x is represented as a directed, generalized random graph [Newman et al. 2001] whose topological characteristics are represented by the degree distribution $\{p(d_x)\}$ and by the probability that a randomly chosen degree d_x account has an outgoing contact pointing to a degree d'_x account $\{p(d'_x|d_x)\}$.

To characterize the mean number of contacts up to distance t that can potentially receive the information, we start by considering a randomly chosen degree $d_x = (i_x, o_x)$ account in OSN x; Equation 1 describes $c_{d_x,d'_x}^{(t)}$ that is the mean number of degree d'_x contacts t hops away in the *same* OSN that receive the information. It is a recursive definition that is based on the main assumption that the network of contacts is locally tree-like[1].

[1]Actual networks of contacts are not tree-like but in Section 4 we show that for small-to-moderate values of t the model predictions are in agreement with simulations hence making this assumption acceptable.

Table I. Paper notation

Parameter	Description
X	Number of interconnected OSNs.
$\mathcal{X}$	Set of indexes to identify OSNs.
	Topology parameters
d_x	Joint in-out degree of a randomly chosen account in OSN x.
$p(d_x)$	Fraction of degree d_x accounts.
$p(d'_x\|d_x)$	Probability that a randomly chosen degree d_x account has an outgoing arc pointing to a degree d'_x account.
$b(d_x, d_y)$	Joint probability that a randomly chosen account in OSN x has degree d_x and it is also a degree d_y account in OSN y.
	User-related parameters
$f_{d_x}^{(t)}$	Probability a degree d_x account in OSN x forwards the information whose age is t.
$q_{d_x}^{(t)}$	Probability a degree d_x account in OSN x selects a contact to forward the information whose age is t.
α_{d_x}	Fraction of active degree d_x accounts in OSN x.
β_{d_x,d_y}	Probability that a degree d_x user in OSN x takes a forwarding action for the information using his/her account in OSN y wherein degree is d_y.
	Model description
$c_{d_x,d'_x}^{(t)}$	Mean number of degree d'_x contacts t hops away that can potentially receive the information originating from a randomly chosen degree d_x account in OSN x.
$r_{d_x,d_y}^{(t)}$	Mean number of degree d_y contacts in OSN y that are t hops away and that can potentially receive the information originating from a randomly chosen degree d_x account in OSN x.

The base case for $t = 1$ represents the sum of o_x i.i.d. Bernoulli variables whose parameter is equal to $p(d'_x|d_x)$, i.e., the probability that a randomly chosen degree d_x account has an outgoing arc pointing to a degree d'_x account.

$$c_{d_x,d'_x}^{(t)} = \begin{cases} o_x \cdot p(d'_x|d_x) & t = 1 \\ \sum_{d''_x} c_{d_x,d''_x}^{(t-1)} \cdot c_{d''_x,d'_x}^{(1)} & t > 1 \end{cases} \tag{1}$$

To consider diffusion in other OSNs, i.e., to generalize Equation 1 to the OSI scenario, we assume we randomly select a degree d_x account that is an *island* in OSN x. Equation 2 defines $r_{d_x,d_y}^{(t)}$ that describes the mean number of contacts t steps away whose degree is equal to d_y in OSN y that can potentially receive the information. For $t = 1$ the mean number of contacts in OSN y is equal to $c_{d_x,d_y}^{(t)}$ if $x = y$ and it is equal to 0 otherwise since we are focusing on an island (the symbol $\delta_{x,y}$ denotes the Kronecker delta).

For larger values of t we must account for the number of contacts in OSN w whose degree is d_w after $t - 1$ steps. A fraction of such contacts is a bridge towards OSN y whose degree is equal to d'_y (this fraction is given by the ratio $\frac{b(d_w,d'_y)}{p(d_w)}$); the contacts potentially reachable in one step whose degree is equal to d_y are then considered by the function composition. The overall number is then described by summing over all possible values of w, d_w, and d'_y.

$$r_{d_x,d_y}^{(t)} = \begin{cases} \delta_{x,y} c_{d_x,d_y}^{(t)} & t = 1 \\ \sum_{\substack{w \in \mathcal{X} \\ d_w, d'_y}} r_{d_x,d_w}^{(t-1)} \cdot \frac{b(d_w,d'_y)}{p(d_w)} c_{d'_y,d_y}^{(1)} & t > 1 \end{cases} \tag{2}$$

Please note that Equation 2 simplifies to Equation 1 when $X = 1$ since in this case we have that $b(d_x, d_y) = p(d_x)$ when x and y coincide and 0 otherwise.

The overall number of potentially reachable contacts from a degree d_x account that are t hops away can be obtained by summing all contributions from Equation 2 yielding

$$r_{d_x}^{(t)} = \sum_{\substack{y \in \mathcal{X} \\ d_y}} r_{d_x,d_y}^{(t)}. \tag{3}$$

We now relax the assumption that the information origin is an island in OSN x and we consider the contribution the origin can give to cross OSN boundaries since the very beginning of the diffusion process. To this end, we complete the description of the potential number of reachable contacts of a degree d_x account that are t hops away as the weighted sum of contributions from Equation 3, i.e.,

$$R_{d_x}^{(t)} = \sum_{\substack{y \in \mathcal{X} \\ d_y}} b(d_x, d_y) r_{d_y}^{(t)}. \tag{4}$$

The final step of our model development allows one to derive the overall mean potential number of contacts up to a maximum distance T receiving the information originating from any account in OSN x. To this end, we define Equation 5 that is obtained by combining all contributions from Equation 4 yielding

$$\overline{p}_{x,T} = \sum_{d_x} p(d_x) \sum_{t=1}^{T} R_{d_x}^{(t)}. \tag{5}$$

The $\overline{p}_{x,T}$ value represents the mean overall number of contacts in the OSI scenario that can structurally receive the information that spreads through the contact relationships among accounts.

3.2. Actual propagation

In the previous section we derived the mean number of contacts that can potentially receive the information originating from a random account in OSN x by properly combining instances of Equation 2. Equation 6 is a refinement of Equation 2 that describes the mean number of contacts t steps away in OSN y that *actually* receive the information originating from a degree d_x account in OSN x.

$$a_{d_x,d_y}^{(t)} = \begin{cases} \delta_{x,y} f_{d_x}^{(0)} q_{d_x}^{(0)} c_{d_x,d_y}^{(t)} & t = 1 \\ \sum\limits_{\substack{w \in \mathcal{X} \\ d_w, d'_y}} \alpha_{d_w} f_{d_w}^{(t-1)} a_{d_x,d_w}^{(t-1)} \dfrac{\beta_{d_w,d'_y} b(d_w, d'_y)}{p(d_w)} q_{d'_y}^{(t-1)} c_{d'_y,d_y}^{(1)} & t > 1 \end{cases} \tag{6}$$

To derive it we considered the:

— forwarding mechanism inside a OSN. We used the forwarding probabilities $\{f_{d_x}^{(t)}\}$ to represent the possibility accounts discard the information and contact selection probabilities $\{q_{d_x}^{(t)}\}$ to represent a random selection of neighbors of forwarding accounts that will actually receive the information. Contact selection probability $q_{d_x}^{(t)}$ can be understood as the fraction of contacts of a degree d_x account that will receive the information whose age is t where ages are represented as the number of hops from the information origin.

— users activity index. To account for the activity index of users we observe that forwarding takes place only when accounts are actually used therefore:

— forwarding occurs only if a degree d_w account in OSN w is active (the probability of this event is equal to α_{d_w});
— with the complementary probability $1 - \alpha_{d_w}$ the number contacts receiving the information is equal to zero.
— cross-posting propensity. Information can cross the OSN boundaries only if two conditions occur: a user is a bridge (this is represented by the probability $\frac{b(d_w,d'_y)}{p(d_w)}$) *and* a user decides to forward the information to outgoing contacts of his/her account in another OSN. This latter condition is represented in Equation 6 by the crossing propensity β_{d_w,d'_y}, i.e., the probability that a degree d_w account in OSN w takes a forwarding action for the information using his/her account in OSN y wherein the degree is d'_y.

Please note that Equation 6 simplifies to Equation 2 when $\forall x, y \in \mathcal{X}, \forall d_x, d_y, t : q^{(t)}_{d_x} = f^{(t)}_{d_x} = \alpha_{d_x} = \beta_{d_x,d_y} = 1$.

The overall mean number of actually reachable contacts from a degree d_x account that are t hops away can be obtained by summing all contributions from Equation 6 yielding

$$a^{(t)}_{d_x} = \sum_{\substack{y \in \mathcal{X} \\ d_y}} a^{(t)}_{d_x,d_y}. \tag{7}$$

Since a degree d_x account (the information origin) can be a bridge we must take into account the contribution the origin can give to cross OSN boundaries since the beginning of the diffusion process. To this end, we complete the description of the actual number of reachable contacts of a degree d_x account that are t hops away as the weighted sum of contributions from Equation 7, i.e.,

$$A^{(t)}_{d_x} = \sum_{\substack{y \in \mathcal{X} \\ d_y}} b(d_x, d_y)\beta_{d_x,d_y} a^{(t)}_{d_y}, \tag{8}$$

The final step of our model development is the definition of the overall actual number of contacts up to a maximum distance T receiving the information originating from any account in OSN x. To this end, we define Equation 9 that is obtained by combining all contributions from Equation 8 yielding

$$\overline{a}_{x,T} = \sum_{d_x} p(d_x) \sum_{t=1}^{T} A^{(t)}_{d_x} \tag{9}$$

3.3. Information propagation efficiency

The model analysis requires the definition of some measures to quantify the impact of the OSI parameters on the efficiency of information propagation. To this end, from Equations 5 and 9 we consider the *information propagation efficiency* as the ratio

$$s_{x,T} = \frac{\overline{a}_{x,T}}{\overline{p}_{x,T}}. \tag{10}$$

This index is actually the relative size of the information propagation and represents a measure of efficiency; the values of $s_{x,T}$ are in the range $[0, 1]$ where $s_{x,T} = 1$ means all accounts potentially reachable from the information origin have actually received it.

4. RESULTS

The following section presents a discussion on important issues in model validation as well as the real-world data we exploited to derive the model parameters that define our ref-

Table II. Snapshots relevant statistics

OSN	Size	Number of bridges between OSN pairs				
		YouTube	LiveJournal	MySpace	Twitter	Google+
YouTube	209,851	NA	23	142	104	319
LiveJournal	147,100	23	NA	27	84	135
MySpace	1,362,128	142	27	NA	100	59
Twitter	1,378,470	104	84	100	NA	1,264
Google+	425,775	319	135	59	1,264	NA

erence scenario. Finally, we discuss some examples of model exploitation to gain a better understanding on which factors influence information spreading in an OSI scenario.

4.1. Issues in model validation

Model validation should consider two features: realism and correctness. As for the model realism, ideally validation should be conducted by mining real-world data to obtain values for the topology of each OSN, the bridge interconnection among them, as well as the user-related parameters. Furthermore, once fed to the model solution algorithm, the predicted information propagation should be compared to the propagation actually observed in real-world data. To fully characterize the model realism this kind of real-world data should be collected for different types of information since, as discussed in Section 2.2, information content is one of the most influential parameters in final information diffusion.

Unfortunately, to the best of our knowledge, logs of information cascades over a set of interconnected OSNs are not available due to the inherent complexity of tracking down simultaneous spreading over multiple media of an information whose format may adapt to the OSN (we actually believe this is still an open problem in OSN analysis and measurement). Indeed, most of the publicly accessible data are either information cascades on a single OSN, or snapshots of contact networks of an isolated OSN, or bridge information (without internal OSN structure) in an OSI scenario. This is not surprising since simultaneously collecting *all* these information by means of crawlers is difficult and time consuming.

Nevertheless, data collected in [Buccafurri et al. 2013] can be exploited to derive all topology-related parameters for our model validation. This leaves us with an arbitrary choice for the user-related parameters; although model realism is reduced we are still able to evaluate our model correctness in closer to reality scenarios. Once correctness is ascertained, our model predictions can be exploited to perform a what-if analysis to gain a better understanding on which factors influence information spreading in an OSI scenario.

Model validation (correctness) is carried out by means of simulations run on the OSI scenario taken from [Buccafurri et al. 2013] and composed of a set of OSN snapshots as well as their bridge interconnections. Validation through comparison against simulation results is the common and inevitable choice of all previous works, e.g., [Li et al. 2015; Yagan et al. 2013].

4.2. Dataset description

Validation of our model relies on the availability of real snapshots of OSN topologies and their bridge interactions to run simulations on them: data collected in [Buccafurri et al. 2013] include both information required for our model validation. In that paper, the authors crawled five OSNs, namely Twitter, YouTube, MySpace, LiveJournal, and Google+. They experimented with several techniques to also crawl the bridge interconnection among OSNs and reported results on the distribution of bridges. In particular, the log files[2] after a bit of post-processing provide the possibility to extract all model parameters related to OSN topologies and bridges distribution as defined in Section 2.

[2]available at http://www.ursino.unirc.it/bridges.html

Table III. Parameters of the reference scenario

OSN	f_x	q_x	α_x	Cross-posting propensities $\beta_{x,y}$				
				YouTube	LiveJournal	MySpace	Twitter	Google+
YouTube	0.5	0.75	0.5	1	0.25	0.5	1	1
LiveJournal	0.25	0.5	0.25	0.25	1	0.5	1	1
MySpace	0.5	0.5	0.5	0.5	0.5	1	1	1
Twitter	0.75	1	0.75	0.25	0.25	0.25	1	0.5
Google+	0.75	1	0.75	0.25	0.25	0.25	0.5	1

Table II shows the size of each snapshot as well as the number of bridges between each pair of OSNs. Since the analysis in [Buccafurri et al. 2013] focused on the distribution of bridges little attention has been paid to coverage of each OSN. As a consequence, the snapshots representing the considered OSNs *are small*; indeed, it is well known that graph size does affect the computation of the average overall number of nodes reachable from a randomly chosen vertex up to distance T. For instance, in the case of entirely random undirected graphs with arbitrary degree distribution one can express this quantity as $1+\sum_{t=1}^{T} z_t$ where $z_t = z_1 \cdot (\frac{z_2}{z_1})^{t-1}$, and where z_1 and z_2 are the average numbers of first and second-nearest neighbors, respectively (Section II.F in [Newman et al. 2001]). Of course, this formula holds only when this quantity is not close (or does not exceed) the considered graph size. In the case of directed, correlated, and interconnected graphs considered in this paper, formulas are more complex (Equations 1-5) but the effect of graph finite (and small) sizes is similar. The OSN graphs used for the model validation are small meaning that their size and degree distributions are such that the equations developed in Section 3 are valid only for very small values of T, i.e., T=2,3.

To cope with the limited size of the snapshots we devised a graph magnification operation (specified by Algorithm 1 in the paper Appendix that increases the size of the graph (the number of nodes) by the magnification factor k. It works by creating k replicas of each node in the original graph; these replica compose the set of nodes V_k. For each node in V_k the same in/out degree is set. After that, each arc of the original graph is analyzed and k replicas are added to E_k that randomly connect nodes in V_k with identical degrees. It can be easily observed that the operation implemented by Algorithm 1 preserves all the topological characteristics of the original OSN contact networks that have been exploited in the model development, i.e., $p(d_x)$ and $p(d'_x|d_x)$, while yielding arbitrarily larger graphs. A similar replication approach is carried out for magnifying the number of bridges to preserve the original $b(d_x, d_y)$ probabilities.

4.3. Reference scenario

The dataset we use allows us to define all the model input parameters related to the OSI topology (see Table I); these parameters are those required to compute the values of Equation 5. To compute values of Equation 9 we also need to define values for user-related model parameters. To this end, we defined a reference scenario whose settings are summarized in Table III. We arbitrarily assigned values to $f_{d_x}^{(t)}$, $q_{d_x}^{(t)}$, α_{d_x}, and β_{d_x,d_y}; to simplify the analysis and the management of model parameters we chose to drop dependencies on both account degrees and information age except for the forwarding probability of origins that is always equal to 1, i.e., $\forall x \in \mathcal{X}, \forall d_x, f_{d_x}^{(0)} = 1$. Nevertheless, we retained system heterogeneity in the definition of all model parameters.

In particular, we assumed that:

— information interest is low in LiveJournal, medium in YouTube and MySpace, and high in Twitter and Google+ (column f_x of Table III represents forwarding probabilities);

— the size of the outgoing contacts receiving the information is low in both LiveJournal and MySpace, medium in YouTube, and maximum in both Twitter and Google+ (column q_x of Table III represents contact selection probabilities);
— user activity in LiveJournal is low while it is medium in YouTube and MySpace, and high for Twitter and Google+ (column α_x of Table III represents user activity).

As for the cross-posting propensities, we chose to tightly couple all OSNs to both Twitter and Google+ and to loosely couple all other OSN pairs. Clearly, within OSN x propensity is maximum, i.e., $\beta_{x,x} = 1$.

Although we defined all parameters for all five OSNs, our reference scenario includes only YouTube, LiveJournal, and Twitter (model parameter $X = 3$); furthermore, the maximum age of information (parameter T in Equations 5 and 9) is set to three hops from the origin. These choices have been forced by the high computational complexity of simulations required for validation. Indeed, simulations were feasible for large magnification factors only for subsets of three out of five OSNs and for rather small values of T. Please note that the complexity of the model solution from parameter T is linear while it is exponential for simulations.

4.4. Model validation

We developed a simulator to represent information propagation on OSN snapshots obtained from [Buccafurri et al. 2013] up to a maximum information age T. To this end, the simulator activity is organized as follows:

— magnification of OSN snapshots using a magnification factor k;
— allocation and initialization of data structures representing X OSNs and bridge relationships among them;
— reproduction of the information propagation from each account n of each OSN x. This activity is simulated by considering the starting account n as the root of a probabilistic breadth-first visit of the interconnected OSNs whose depth is equal to T. At the t^{th} step of the visit a degree d_x account in OSN x decides to forward or to drop the information according to its activity α_{d_x} and to the forwarding probabilities $f_{d_x}^{(t)}$. Furthermore, if the account propagates the information it does so by using contact selection probability $q_{d_x}^{(t)}$ to select receiving outgoing contacts. Finally, if the account is also a bridge towards a degree d_y account in OSN y then propensity β_{d_x,d_y} is used to decide to cross-post the information to OSN y.

Each information propagation operation records the overall number of contacts that received the information. The average of this quantity is computed for each OSN x and is denoted as $\hat{a}_{x,T}$. The simulations are run in 30 independent trials to compute 95% confidence intervals to be compared against Equation 9 computed using as inputs distributions $p(d_x)$, $p(d'_x|d_x)$, and $b(d_x, d_y)$ measured from snapshots in [Buccafurri et al. 2013].

Validation has been carried out in two cases:

— first, by setting all user-related parameters of our model ($\{f_{d_x}^{(t)}\}$, $\{\alpha_{d_x}\}$, $\{q_{d_x}^{(t)}\}$, and $\{\beta_{d_x,d_y}\}$) to 1 to compare $\hat{p}_{x,d}$ against Equation 5 ($\overline{p}_{x,T}$);
— then, by using parameters that define our reference scenario (summarized in Table III) to compare $\hat{a}_{x,T}$ against Equation 9 ($\overline{a}_{x,T}$).

Figure 1 depicts the absolute relative error (defined as $|\frac{\hat{a}_{x,T}-\overline{a}_{x,T}}{\hat{a}_{x,T}}|$) between the predictions of our model and the same quantities estimated from simulations in the OSI reference scenario composed of YouTube, LiveJournal, and Twitter (other subsets of OSNs provided similar results). It can be noted that simulation results and model predictions are in excellent agreement for the three OSNs we chose as the snapshot size increases. This confirms

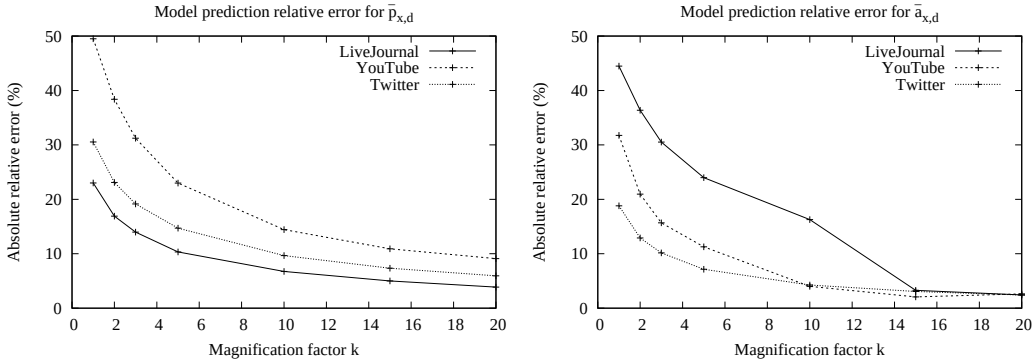

Fig. 1. Model predictions relative error for increasing values of graph magnification factor k.

Table IV. Average actual number of contacts and information efficiency for increasing complexity of the OSI scenario

OSN	YouTube	LiveJournal	Twitter	all three	three + MySpace	four + Google+
YouTube	3.24/0.132	3.26/0.088	28.40/0.278	28.57/0.237	85.36/0.006	21,638.54/0.260
LiveJournal	7.78/0.002	7.78/0.002	78.40/0.015	78.41/0.015	101.64/0.009	1,620.20/0.091
Twitter	5,131.65/0.329	5,132.97/0.328	5,131.38/0.329	5,133.24/0.328	5,140.67/0.270	9,686.67/0.201

that our model is well suited for large scale OSNs and is accurate enough to provide trustful predictions on the characterization of information propagation in complex OSI scenarios. The paper Appendix further discusses how the graph magnification affects structural properties other than degree distribution and degree correlations, and how the network size itself affects the accuracy of model predictions.

Please note that we limited our simulations to only subsets of three out of five available OSNs, i.e., $X = 3$, for managing computational complexity of simulations: indeed, only *one* simulation experiment for magnification factor $k = 20$ took an average of 5,400s while the model solution required only 135s on the same computer. Since the time complexity of simulations is exponential with T and X this means that experiments for larger scenarios take days of CPU time to complete.

4.5. Model exploitation

This section shows how to exploit the model predictions to gain a better understanding on which factors influence information spreading in an OSI scenario.

4.5.1. Information propagation in isolated OSNs. Here we take a closer look to the model predictions whose accuracy with respect to simulations we discussed in the previous section. To this end, Table IV shows the impact of the complexity of the OSI scenario on the average overall number of actual contacts that received the information ($\overline{a}_{x,T}$ in Equation 9) and the information efficiency ($s_{x,T}$ in Equation 10). Gray-shaded cells in the first three columns refer to an OSN in isolation and from them it can be noted that information spreads very little and rather inefficiently on both YouTube and LiveJournal but for different reasons:

— for YouTube, this is due to the topological characteristics of this OSN as captured by the snapshots we used; indeed, the average overall number of potential contacts can be obtained from Equation 10 as ($\overline{p}_{x,T} = \frac{\overline{a}_{x,T}}{s_{x,T}}$) and it is equal to only 24.63 for YouTube;
— for LiveJournal, $\overline{p}_{lj,T} = 3,204.86$ (a much higher value) but information spreads very little because of the particularly low values we used for f_{lj}, α_{lj}, and q_{lj} in the reference scenario described in Table III.

On the contrary, Twitter has both large values of $\overline{p}_{tw,T} = 15,574.85$ and displays high information efficiency thanks to high values for $f_{tw} = \alpha_{tw} = 0.75$ and to the broadcast nature of tweets represented by $q_{tw} = 1$.

4.5.2. Information propagation in OSN pairs. Table IV also shows how information spreads when an OSN is part of a more complex OSI scenario. In particular:

— propagation of information originating in YouTube (first row) is virtually unaffected by the coupling with LiveJournal. About 90% of the bridges from YouTube to LiveJournal have at least one outgoing contact in LiveJournal (average number of outgoing contacts is equal to 185.261) but have low crossing propensity ($\beta_{d_{yo},d_{lj}}$ is only 0.25 in the reference scenario); furthermore, analysis of information propagation in LiveJournal in isolation highlighted the limiting effect of the values of f_{lj}, α_{lj}, and q_{lj} we chose in the reference scenario. Coupling with Twitter shows limited benefit for information originating in YouTube despite the high value of $\overline{p}_{tw,T}$ observed for Twitter in isolation. Indeed, this can be explained by noting that about 75% of the bridges from YouTube to Twitter have 0 outgoing contacts in Twitter although the average number of outgoing contacts in Twitter was equal to 400.472 in the snapshots we considered;
— information originating in LiveJournal cannot use YouTube as a mean to reach more contacts since bridges from LiveJournal to YouTube all have 0 outgoing contacts in YouTube in the snapshots we consider: the final effect is that information cross-posted from LiveJournal to YouTube does not spread any further in this additional OSN. Moreover, also in this case coupling with Twitter shows limited benefit for information propagation: indeed, about 65% of the bridges from LiveJournal to Twitter have 0 outgoing contacts in Twitter although the average number of outgoing contacts in Twitter was equal to 215.69;
— when information originates in Twitter it gains no benefit from interconnecting to either OSNs despite the non-zero connectivity of bridges from Twitter to YouTube (about 75% of bridges in YouTube have at least one outgoing contact and the average is equal to 112.245) and from Twitter to LiveJournal (almost all bridges in LiveJournal have at least one outgoing contact and the average is equal to 347.214); in this case, the limiting factor is the rather small cross-posting propensity in our reference scenario of Twitter bridges toward both YouTube and LiveJournal, i.e., $\beta_{tw,yo} = \beta_{tw,lj} = 0.25$.

4.5.3. Impact of the OSI size. The shaded column in Table IV shows the information propagation size in the complete reference scenario composed of YouTube, LiveJournal, and Twitter. It can be noted that results are (slightly) greater than the isolated and pair cases with results that are very close to those obtained by the interaction with Twitter only. Of course, the explanation we provided to justify results in the OSN pairs case hold in the complete scenario as well.

Table IV is completed with the last columns where MySpace and Google+ are also included in the model analysis. It can be noted that all three reference OSNs observe a marked increase in the values of $\overline{a}_{x,T}$ since Google+ is by far the OSN with the largest value of $\overline{a}_{x,T}$ and all other OSNs have maximum propensity to cross-post the information towards it, i.e., the values of $\beta_{*,go}$ are equal to 1 for almost all OSNs. Furthermore, bridges from all OSNs to Google+ have often non-zero outgoing contacts in Google+ in the snapshots we considered.

4.5.4. Impact of bridge degree correlations. In the previous section we observed the key role of bridges in propagating the information in an OSI scenario. In this section we further analyze their role by first evaluating the values of $\overline{a}_{x,T}$ in two cases: first, we add new bridges to the reference scenario and then we modify the degree distribution of the existing ones.

— For the first part of this analysis we add b_{add} new bridges to each subset of the X snapshots we consider. We choose accounts in each OSN such that their out-degree is equal to d_{add}

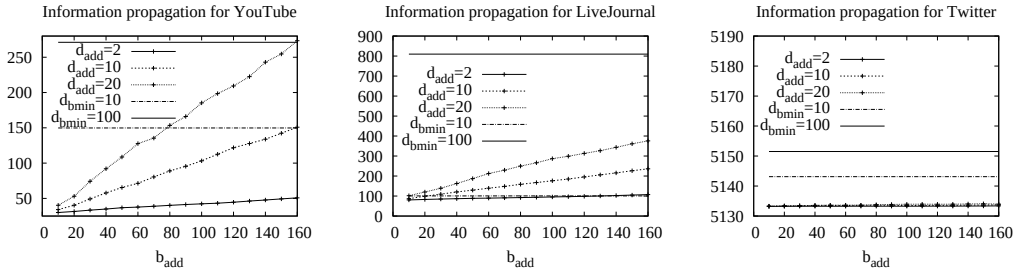

Fig. 2. Average number of accounts receiving the information ($\overline{a}_{x,T}$) as a function of the number of additional bridges (b_{add}).

Table V. Average actual number of contacts and information efficiency for increasing bridge degrees

OSN	all three	$d_{bmin} = 10$	$d_{bmin} = 100$
YouTube	28.57/0.237	149.77/0.212	271.17/0.238
LiveJournal	78.41/0.015	100.79/0.020	809.74/0.047
Twitter	5,133.24/0.328	5,143.12/0.322	5,151.52/0.313

connections to tune overall bridge connectivity. We then compute probabilities $b(d_x, d_y)$ to solve our model. Please note, that this transformation affects the distribution of bridges between OSNs and does not preserve the total fraction of contacts in each OSN that is a bridge, i.e., $\forall x, y \in \mathcal{X}, \sum_{d_x,d_y} b(d_x, d_y)$.

Figure 2 shows the values of $\overline{a}_{x,T}$ for increasing values of b_{add} and for different values of d_{add}. It can be noted that information propagation linearly increases as the number and out-degree of additional bridges increases. This phenomenon is marked for YouTube and LiveJournal while it is marginal for Twitter that, with the range of values we considered, almost does not benefit from interconnection to other OSNs.

— In the previous section, we observed that low out-degree bridges in one OSN limit the actual diffusion of information in the reference OSI scenario. To better understand the effect of this limitation we modify the $b(d_x, d_y)$ probabilities as follows: for each bridge between OSNs x and y in the original snapshots we select two randomly chosen contacts (one in each OSN) such that the in-degree and the out-degree of both exceeds a given threshold d_{bmin}. Starting from the modified snapshots we compute the new $b(d_x, d_y)$ probabilities to solve our model and compute $\overline{a}_{x,T}$ and $s_{x,T}$ for all OSNs. Please note that this transformation only affects the distribution of bridges between OSNs but preserves the total fraction of contacts in each OSN that is a bridge, i.e., $\forall x, y \in \mathcal{X}, \sum_{d_x,d_y} b(d_x, d_y)$ is unchanged.

Table V shows results for the original scenario (the gray shaded column is taken from Table IV to ease comparison) and the results obtained for increasing values of the degree threshold d_{bmin}. It can be noted that information originating in both YouTube and LiveJournal propagates now to a larger number of contacts as the degree of their mutual bridges increases. In particular, information originating in LiveJournal can now reach an increased number of contacts in YouTube; moreover, despite the low cross-posting propensity of YouTube bridges toward LiveJournal it is now possible for information originating in YouTube to reach a larger number of LiveJournal contacts. Diffusion of information originating in Twitter is only slightly affected because cross-posting propensity remains low in the reference scenario while degree of bridges in the original snapshots was already high: the average number of outgoing contacts of bridges towards YouTube was equal to 112.245 and it was equal to 347.214 for bridges towards LiveJournal.

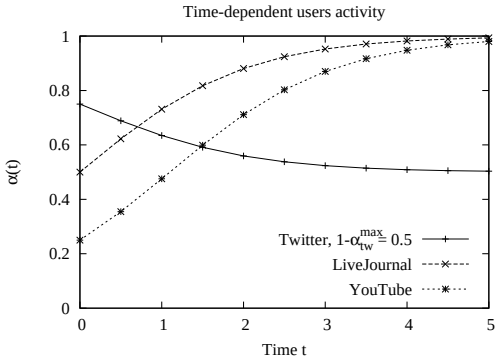

Fig. 3. Time-dependent users activity.

We also plotted the values of Table V in the graphs of Figure 2. For YouTube, we note that adding 80 new bridges whose out-degree is equal to 20 is as effective as considering existing bridges with minimum degree $d_{bmin} = 10$. It takes a lot more bridges to obtain the same effect when $d_{bmin} = 100$. It can also be noted that for LiveJournal and Twitter increasing the degree of existing bridges is much more effective than adding new low degree bridges to obtain a higher number of accounts that receive the information.

4.5.5. Impact of OSN evolution. In this section we consider the growth and decline of OSNs in time and analyze how the information propagation can be affected by these phenomena. We consider the scenario where bridge degrees exceed $d_{bmin} = 100$ and we represent time-dependent OSN growth and decline by increasing (decreasing, respectively) the fraction of active users in each OSN. We let user activities depend on time where we denote as $\alpha_x(t)$ the time-varying parameter for users in OSN x. To represent the temporal evolution of user activities we assume Twitter is declining in time while Youtube and LiveJournal enjoy explosive growth; we refer to [Ribeiro 2014] and we set all time dependencies as exponential laws of the kind

$$g_x(t) = \frac{k_x^{max}}{1 + e^{-(t-h_x)}}. \tag{11}$$

In this case, function $g_x(t)$ represents $\alpha_x(t)$ where parameter k_x^{max} is equal to α_x^{max} (the asymptotic value for $\alpha_x(t)$); parameter h_x is set to start at time 0 with $\alpha_x(t) = \alpha_x$ as defined in the reference scenario (Table III) and to asymptotically reach α_x^{max} for increasing user activities and $1 - \alpha_x^{max}$ for decreasing activities (we consider $\alpha_{yo}^{max} = \alpha_{lj}^{max} = 1$ and $1-\alpha_{tw}^{max} = 0.4, 0.5, 0.6$). Figure 3 displays the time evolution we consider for $1-\alpha_{tw}^{max} = 0.5$.

Figure 4 shows that for YouTube and Twitter the decaying activity of Twitter users has a negative impact on the number of contacts reached by the information originating in them. The larger the decay the lower the number of informed accounts. We already observed that information originating in Twitter actually propagates mostly inside it therefore it is rather straightforward to observe a reduction in the values of $\overline{a}_{tw,T}$. For YouTube, the increase of the number of internal accounts reached thanks to OSN growth is counterbalanced by the more marked decrease of Twitter accounts receiving the information.

For LiveJournal the phenomenon is a little less intuitive, instead. On one hand, the number of contacts inside LiveJournal that receive the information increases thanks to the time increasing $\alpha_{lj}(t)$; on the other hand, the number of Twitter accounts receiving the information from it decreases. Depending on the intensity of decaying, i.e., the values of $1 - \alpha_{tw}^{max}$, the two contributions either balance or dominate each other. When decaying is marked ($1 - \alpha_{tw}^{max} = 0.4$) LiveJournal information spreads less in time while for lighter

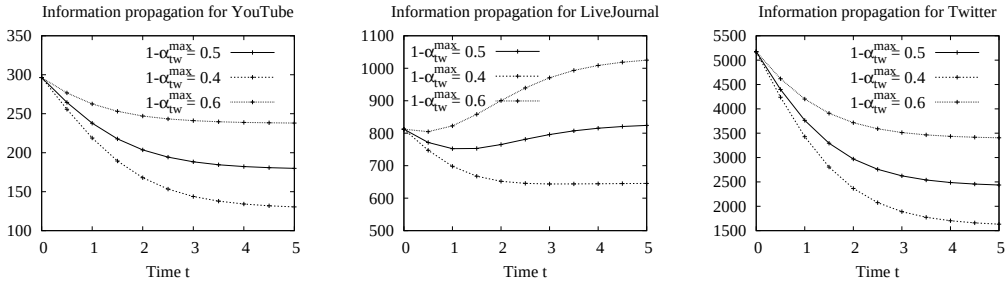

Fig. 4. Model predictions ($\overline{a}_{x,T}$) for OSN evolution in time.

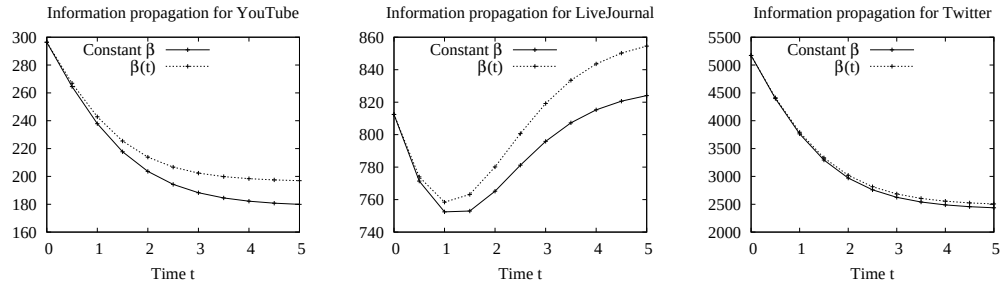

Fig. 5. Model predictions ($\overline{a}_{x,T}$) for time-dependent cross-posting propensities and user activities for $1 - \alpha_{tw}^{max} = 0.5$.

decaying ($1 - \alpha_{tw}^{max} = 0.6$) propagation increases. The intermediate case ($1 - \alpha_{tw}^{max} = 0.5$) displays a non-monotonic time behavior where the initial evolution leads to a decrease of $\overline{a}_{tw,T}$ while as time passes by the average overall number of reached contacts starts to increase again.

4.5.6. Impact of cross-posting propensity. In this section we consider how the cross-posting propensity of OSN users [Reza Farahbakhsh and Crespi 2015; Ottoni et al.] affects the efficiency of information propagation. To this end, we again inspire to [Ribeiro 2014] to set the values of $\beta_{x,y}$ parameters of our model. In particular, we consider the reference scenario where all cross-posting propensities lesser than 1 evolve in time following an exponential law as in Equation 11; in this case, function $g_x(t)$ represents $\beta_{x,y}(t)$ where parameter k_x^{max} is equal to 1 for all cases of the reference scenario (Table III) where $\beta_{x,y}$ values are lesser than 1, i.e., $\beta_{yo,lj}, \beta_{lj,yo}, \beta_{tw,yo}$, and $\beta_{tw,lj}$. Parameter h_x is set to start at time 0 with $\beta_{x,y}(t) = \beta_{x,y}$ as defined in the reference scenario and to asymptotically reach $\beta_{dx,dy}^{max}$. Figure 5 shows that increasing cross-posting propensities has a limited impact on the diffusion of information in the OSI scenario we considered. Indeed, the main forces driving the spread of information throughout OSNs are both the total fraction of contacts in each OSN that is a bridge, i.e., $\forall x, y \in \mathcal{X}, \sum_{d_x,d_y} b(d_x, d_y)$ and the degree distribution of bridges as already observed.

5. RELATED WORK

Information propagation in a single isolated OSN is well-understood as testified by a large body of previous work, e.g., [Zhao et al. 2012; Bakshy et al. ; Yang and Leskovec ; Kumar et al. 2006], and by two recent surveys on this topic [Hu et al. 2015; Guille et al. 2013]. Modeling and analysis of information spreading in multiple OSN is still in its infancy although a

substantial amount of research has been carried out on the more general and abstract topic of diffusion processes on multilayer networks [Salehi et al. 2015].

A few papers focused on information spreading in multiple OSNs. In particular, we consider [Li et al. 2015] where the authors propose an information diffusion model on multiplex networks with two layers composed of the same number of nodes representing the same group of users. The paper makes a lot of simplifying assumptions to be able to derive analytical results. It neglects correlations in the OSN topologies and in bridge interconnections. It does not include any of our user-related model parameters and it does not provide validation on real snapshots.

A second work we believe is related to ours is [Yagan et al. 2013] where the authors derive analytical results for a system composed of an overlay and a physical network the information can use to propagate. The limitations of this work are the same as [Li et al. 2015].

Validation of our model results requires the collection and exploitation of real world data mined simultaneously from multiple OSNs. This is an arena where several problems related to concepts definition and measurement approaches can still be considered as open. Therefore, only partial real world data are available to modelers, i.e., topological information of the OSI scenario as reported in [Buccafurri et al. 2013]. The authors of [Su 2014] present a clear picture of the current state-of-the art of studies aiming at mining data on multiple OSNs.

6. CONCLUSIONS AND FUTURE WORK

The main contribution of this paper is the development of a tractable mathematical model for the analysis of information propagation across multiple OSNs. This problem is highly complex because of many factors: structural heterogeneities of each OSN, bridges that interconnect them, and heterogeneities associated with users such as their activity patterns, interests, and propensities.

In particular, we provided a directed generalized random graph-based model to analyze the information propagation in a complex scenario where users may subscribe to multiple OSNs becoming bridges that can cross-post information to additional accounts. We modeled the correlations that are often found in the topological characteristics of OSNs as well as the co-location of bridges in multiple OSNs. We also represented the information diffusion process by taking into account user interest in information, information age, forwarding mechanism offered by OSNs, users activity, and cross-posting propensity.

We validated the model predictions against simulations run on real snapshots describing a complex OSI scenario and found excellent agreement for large scale systems. We also exploited the model predictions to get insights on the information propagation process as a function of size and complexity of the OSI scenario, degree distribution of bridges, growth and decline of OSNs in time, and time-varying cross-posting users propensity. We observed interesting relationships between degree distribution and overall number of bridges on the actual information propagation as well as non-monotonic behavior when popularity of OSNs evolves in time.

The model we developed represents one step forward in the research of information diffusion across multiple OSNs. Some questions remain open and suggest potential directions for further developments. In particular, the model could be extended to incorporate information propagating from multiple originating accounts. This is a difficult problem to cope with and simple mathematical tools are still lacking.

Any modeling efforts has to face the problem of validating predictions against real measures. Such validation requires the collection and exploitation of real world data simultaneously mined from multiple OSNs. This is an arena where several problems related to concepts definition and measurement approaches can still be considered as open.

Efficient algorithms for the numerical model solution are required when the size of the OSI scenario under investigation grows. This calls for both approximation techniques (to reduce the computational and storage complexity of the model solution) and ad-hoc parallel algorithms exploiting available multi-core processors.

Finally, the model could be used to analyze the propagation of *malicious* information, e.g., fake news, in a multiple social network scenario with a focus on the dynamic evolution of both the topological structure of OSNs and their bridge interconnection.

APPENDIX

The magnification algorithm we informally described in Section 4.2 is completely specified in Algorithm 1. It can be easily noted that it preserves both topological characteristics of the original OSN contact networks that have been exploited in the model development, i.e., $\{p(d_x)\}$ and $\{p(d'_x|d_x)\}$, while yielding arbitrarily larger graphs. The input for the algorithm is the contact network $G = (V, E)$ as well as the magnification factor k. The output is a magnified contact network $G_k(V_k, E_k)$.

ALGORITHM 1: Input: $G = (V, E)$, magnification factor k. Output: $G_k(V_k, E_k)$

```
G_k = ∅, E_k = ∅
for all d do
   s_d = ∅
end for
for all v ∈ V do
   d = (i_v, o_v)
   for j = 0 to k − 1 do
      v' = j · |V| + v
      s_d = s_d ∪ {v'}
      stubs_v' = i_v
      G_k = G_k ∪ {v'}
   end for
end for
for all (v_f, v_t) ∈ E do
   d_t = (i_v_t, o_v_t)
   for j = 0 to k − 1 do
      v'_f = j · |V| + v_f
      v'_t=select random node ∈ s_d_t
      E_k = E_k ∪ {(v'_f, v'_t)}
      stubs_v'_t = stubs_v'_t − 1
      if (stubs_v'_t == 0) then
         s_d_t = s_d_t − {v'_t}
      end if
   end for
end for
```

There is some resemblance between Algorithm 1 and the work in [Faqeeh et al. 2015] where the authors devise a magnification technique (that is called L−cloning) to study the impact of clustering on dynamical processes running on *undirected* networks. Figure 6 depicts the clustering coefficient (CC) [Fagiolo 2007] of magnified directed networks we used in Section 4. It can be noted that Algorithm 1 produces less clustered networks and it thus transforms the magnified graphs into more tree-like structures. This, in turn, affects the accuracy of the model predictions when compared to simulation as shown in graphs of Figure 1 in the paper.

Nevertheless, a lower clustering coefficient is not the only explanation for a higher model accuracy. Indeed, network size alone is responsible for more accurate model predictions, independently of CC values. To verify this, we retrieved other snapshots of the same OSNs we consider in Section 4. We were able to obtain other snapshots for all five OSNs but we could not process the huge Twitter data from [Kwak et al. 2010] [3] and from [Cha et al. 2010][4]

[3]available at http://an.kaist.ac.kr/traces/WWW2010.html
[4]available at http://twitter.mpi-sws.org/

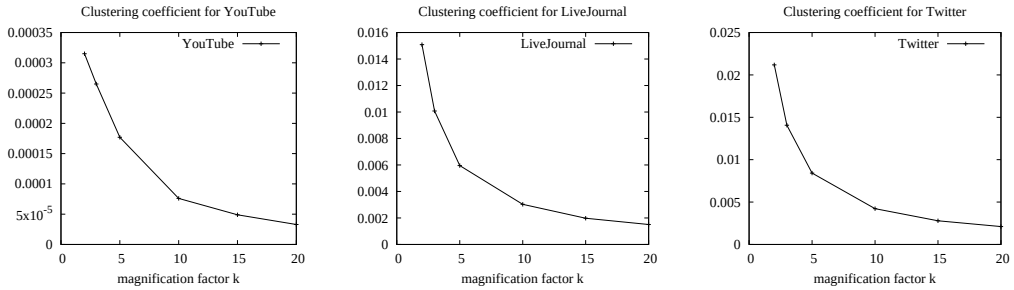

Fig. 6. Clustering coefficient of magnified OSN graphs as a function of magnification factor k.

Table VI. Additional data characteristics

OSN	Network Size (NS) this paper / other source	Clustering Coefficient (CC) this paper / other source	Reference (available at)
YouTube	209,851 / 1,157,827	0.005716 / 0.082344	[Mislove et al. 2007][5]
LiveJournal	147,100 / 4,847,571	0.085915 / 0.163194	[Backstrom et al. 2006][6] [Leskovec et al. 2009]
LiveJournal	147,100 / 5,284,457	0.085915 / 0.159267	[Mislove et al. 2007][7]
Google+	425,775 / 211,187	0.156871 / 0.064898	[Fire et al. 2013][8]
MySpace	1,362,128 / 100,000	0.025233 / 0.128688	[Ahn et al. 2007][9]

due to the limited available computing and storage resources, i.e., an i7 CPU equipped with 4GB RAM. Table VI shows the network size (NS) and the CC for the cases we considered.

We also explored other values for the user related model parameters besides those defined for the reference scenario in Table III; to keep things simple, we selected a scenario where $f = q = \alpha = p$ with $p \in \{0.25, 0.375, 0.5\}$. Table VII shows the absolute relative error (RE) between the model predictions and the simulation results *without any magnification* and for each OSN in isolation. It can be noted that:

— for YouTube the larger snapshot is associated with a *much* larger CC but for some values of the model parameters, i.e., for $p = 0.375$, the model predictions are more accurate;
— a similar observation can be made for the two additional snapshots of LiveJournal: their CC is about twice the one obtained from the snapshots used in this paper but for small values of p the model predictions are more accurate;
— in the case of Google+, the snapshot used in this paper is larger than the only additional one it was possible to retrieve. Furthermore, its CC is more than twice the smaller one. Nevertheless, the RR is lower for all values of p;
— model predictions on MySpace are always more accurate on the larger snapshot (the one used in this paper) when compared to the accuracy obtained on the additional snapshot whose size is just 13 times smaller. It should also be noted that the CC of the smaller snapshot is five times the CC of the larger network.

Please note that we observed similar trends also for f, q, α values of the reference scenario as described in Table III. In particular, all cases but YouTube (in this case the RR is 2% for the smaller snapshot and 18% for the larger one) resulted in higher accuracy of the model predictions when larger snapshots were used to obtain the topology related parameters.

[5]http://socialnetworks.mpi-sws.org/datasets.html
[6]https://snap.stanford.edu/data/soc-LiveJournal1.html
[7]http://socialnetworks.mpi-sws.org/datasets.html
[8]http://proj.ise.bgu.ac.il/sns/googlep.html
[9]https://an.kaist.ac.kr/traces/WWW2007.html

Table VII. Absolute relative error (in %).

OSN	this paper / other source f=q=α=0.25	this paper / other source f=q=α=0.375	this paper / other source f=q=α=0.5
YouTube	2 / 3	7 / 4	5 / 12
LiveJournal	10 / 0.004	0.0001 / 5	6 / 17
LiveJournal	10 / 0.003	0.0001 / 5	6 / 18
Google+	0.0007 / 0.02	7 / 10	24 / 37
MySpace	2 / 3	4 / 24	9 / 90

To conclude, the graph magnification algorithm yields network with increased size and decreased CC: both have an independent effect on the accuracy of the model predictions when compared to simulations on larger snapshots.

REFERENCES

Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. 2007. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*. ACM, New York, NY, USA, 835–844. DOI:http://dx.doi.org/10.1145/1242572.1242685

Konstantin Avrachenkov, Koen De Turck, Dieter Fiems, and Balakrishna Prabhu. Information dissemination processes in directed social networks. In *SOCNET 2014*.

Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*. ACM, New York, NY, USA, 44–54. DOI:http://dx.doi.org/10.1145/1150402.1150412

Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The Role of Social Networks in Information Diffusion. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*.

Francesco Buccafurri, Vincenzo Daniele Foti, Gianluca Lax, Antonino Nocera, and Domenico Ursino. 2013. Bridge analysis in a Social Internetworking Scenario. *Information Sciences* 224, 0 (2013), 1 – 18.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *International AAAI Conference on Web and Social Media*.

Meeyoung Cha, Alan Mislove, Ben Adams, and Krishna P Gummadi. 2008. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*. ACM, 13–18.

Meeyoung Cha, Alan Mislove, and Krishna P Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW 2009*.

Giorgio Fagiolo. 2007. Clustering in complex directed networks. *Physical Review E* 76 (2007), 026107. Issue 2.

Ali Faqeeh, Sergey Melnik, and James P. Gleeson. 2015. Network cloning unfolds the effect of clustering on dynamical processes. *Physical Review E* 91 (2015). Issue 5.

M. Fire, R. Tenenboim, R. Puzis, O. Lesser, L. Rokach, and Y. Elovici. 2013. Computationally Efficient Link Prediction in Variety of Social Networks. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 1 (2013), 10.

Roberto González, Rubén Cuevas, Ángel Cuevas, Reza Farahbakhsh, Reza Motamedi, and Reza Rejaie. Characterization of information propagation in Google+ and its Comparison with Twitter. (????).

Roberto Gonzalez, Ruben Cuevas, Reza Motamedi, Reza Rejaie, and Angel Cuevas. Google+ or Google-?: dissecting the evolution of the new OSN in its first year. In *WWW 2013*.

Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A. Zighed. 2013. Information Diffusion in Online Social Networks: A Survey. *SIGMOD Rec.* 42, 2 (2013), 17–28.

Changjun Hu, Wenwen Xu, and Peng Shi. 2015. Information Diffusion in Online Social Networks: Models, Methods and Applications. In *Web-Age Information Management*, Xiaokui Xiao and Zhenjie Zhang (Eds.). Lecture Notes in Computer Science, Vol. 9391. Springer International Publishing, 65–76. DOI:http://dx.doi.org/10.1007/978-3-319-23531-8_6

Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. 2008. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*. ACM, 19–24.

Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2006. Structure and Evolution of Online Social Networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a Social Network or a News Media?. In *Proceedings of the 19th ACM International Conference on World Wide Web, WWW '10*.

Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. 2009. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics* 6, 1 (2009), 29–123.

Weihua Li, Shaoting Tang, Wenyi Fang, Quantong Guo, Xiao Zhang, and Zhiming Zheng. 2015. How multiple social networks affect user awareness: The information diffusion process in multiplex networks. *Phys. Rev. E* 92 (Oct 2015), 042810. Issue 4. DOI:http://dx.doi.org/10.1103/PhysRevE.92.042810

Han Liu, Atif Nazir, Jinoo Joung, and Chen-Nee Chuah. 2013. Modeling/predicting the evolution trend of osn-based applications. In *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 771–780.

Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. Mining Topic-level Influence in Heterogeneous Networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*.

Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. Measurement and Analysis of Online Social Networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC '07*.

M. E. J. Newman, S. H. Strogatz, and D. J. Watts. 2001. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* 64 (Jul 2001), 026118. Issue 2.

R.a Ottoni, D.L.a Casas, J.P.a Pesce, Jr.a Meira, W., C.b Wilson, A.b Mislove, and V.a Almeida. Of pins and tweets: Investigating how users behave across image-and text-based social networks. In *ICWSM 2014*. 386–395.

Reza Rejaie, Mojtaba Torkjazi, Masoud Valafar, and Walter Willinger. 2010. Sizing up online social networks. *Network, IEEE* 24, 5 (2010), 32–37.

Angel Cuevas Reza Farahbakhsh and Noel Crespi. 2015. Characterization of cross-posting activity for professional users across Major OSNs. In *IEEE/ACM ASONAM, Paris, France, 2015*.

Bruno Ribeiro. 2014. Modeling and Predicting the Growth and Death of Membership-based Websites. In *Proceedings of the 23rd International Conference on World Wide Web*. 653–664.

M. Salehi, R. Sharma, M. Marzolla, M. Magnani, P. Siyari, and D. Montesi. 2015. Spreading Processes in Multilayer Networks. *IEEE Transactions on Network Science and Engineering* 2, 2 (2015), 65–83.

W. C. Su. 2014. Integrating and mining virtual communities across multiple Online Social Networks: Concepts, approaches and challenges. In *Fourth International Conference on Digital Information and Communication Technology and its Applications (DICTAP 2014)*. 199–204.

Mojtaba Torkjazi, Reza Rejaie, and Walter Willinger. 2009. Hot today, gone tomorrow: on the migration of MySpace users. In *Proceedings of the 2nd ACM workshop on Online social networks*. ACM, 43–48.

S. Wen, M. S. Haghighi, C. Chen, Y. Xiang, W. Zhou, and W. Jia. 2015. A Sword with Two Edges: Propagation Studies on Both Positive and Negative Information in Online Social Networks. *IEEE Trans. Comput.* 64, 3 (2015), 640–653.

O. Yagan, Dajun Qian, Junshan Zhang, and D. Cochran. 2013. Conjoining Speeds up Information Diffusion in Overlaying Social-Physical Networks. *Selected Areas in Communications, IEEE Journal on* 31, 6 (2013), 1038–1048.

Jaewon Yang and Jure Leskovec. Modeling Information Diffusion in Implicit Networks. In *Proceedings of the 2010 IEEE International Conference on Data Mining (ICDM '10)*.

Shaozhi Ye and S.Felix Wu. 2010. Measuring Message Propagation and Social Influence on Twitter.com. In *Social Informatics*. LNCS, Vol. 6430.

Jichang Zhao, Junjie Wu, Xu Feng, Hui Xiong, and Ke Xu. 2012. Information propagation in online social networks: a tie-strength perspective. *Knowledge and Information Systems* 32, 3 (2012), 589–608.