Analytical Model for Topology Dependence in Peer-to-Peer Anti-Entropy Spreading

Emre Iskender
Department of Computer Engineering
Koc University
Istanbul, Turkey
eiskender@ku.edu.tr

Mine Çağlar
Department of Mathematics
Koc University
Istanbul, Turkey
mcaglar@ku.edu.tr

Öznur Özkasap
Department of Computer Engineering
Koc University
Istanbul, Turkey
oozkasap@ku.edu.tr

***Abstract*—We examine spreading of epidemics for an anti-entropy algorithm in networks with various P2P (peer-to-peer) overlay topologies. Neighborhood knowledge among peers and information exchange based on proximity are considered. Our analytical model for SI (Susceptible-Infected) epidemics involves equations for calculating the infection probability of each peer in consecutive epidemic rounds as a function of the topology. Using numerical evaluations, we study the effect of graph properties on dissemination as an aspect of real world P2P overlays.**

I. Introduction

Epidemic spreading in a network takes place from infectious nodes to susceptible nodes, and it is modeled as a process in an undirected graph with nodes where every infectious node exchanges information with one of its neighbors. Modeling the spread of epidemics by taking into account the topological and nodes' neighborhood information provides benefits such as predicting the future spreading behavior, developing methods to control epidemics or achieving faster epidemic information dissemination. In prior work, for SIS (Susceptible - Infected - Susceptible) model, different epidemic thresholds are identified in relation to various topological properties of the underlying network [1], [2]. Such properties include average connectivity, connectivity divergence of the topology and maximum eigenvalue of the adjacency matrix. SIS model is applicable in security services in particular to spread of internet worms and e-mail viruses. The epidemic threshold is significant for detecting if the epidemics will spread to the entire network or not.

In this study, we investigate the impact of topology on SI epidemic model, which is suitable for the applications of content dissemination. Topological properties considered for SIS model as well as graph invariants such as degree distribution and eigenvalues are studied as an aspect of real world P2P networks. In P2P content dissemination systems such as BitTorrent [3] and SeCond [4], each peer exchanges information with a group of its neighbors on the overlay. We introduce a model for calculating the infection probabilities of the nodes as a function of the topology through a general adjacency matrix and show our numerical results on various power-law and Erdös-Rényi random topologies.

Epidemic spreading is examined by calculating the infection probabilities of all the nodes in the network for every epidemic round with the pull based anti-entropy algorithm [5]. In the pull approach, when an infectious peer (holding data to be shared) picks a susceptible peer (lacking the specific data) randomly, this triggers data dissemination from infectious peer to the susceptible. Spreading updates are triggered by susceptible peers when they are picked as targets by infectious peers. In contrast to current study, partial membership knowledge among peers and information exchange based on proximity have not been considered in [5].

The paper is organized as follows. In the next section, we state the basic definitions related to epidemic dissemination. The related work is summarized in Section 3. Section 4 gives the details of the proposed model for topology dependence. Numerical results are presented in Section 5. Finally, Section 6 concludes the paper.

II. Principles on Epidemic Spreading

In this section, we give information about the types of epidemic models and define epidemic dissemination approaches.

A. Epidemic Models

1) SI (Susceptible-Infected): In this model, infectious peers are never cured and continue to infect the remaining susceptible peers until the infection is spread among the network. Information dissemination over a network is defined with SI model in [5].

2) SIS (Susceptible-Infected-Susceptible): In this model an infectious peer turns to be a susceptible peer after the cure [6]. But the nodes may become infected again without any restriction.

3) SIR (Susceptible-Infected-Removed): This model is used to represent virus/worm propagation in distributed systems [6]. There are two different proposed models for SIR model: In the first model, each infectious peer is detected and removed from the system. In this model, there exist only infectious and susceptible peers and the population size decreases dynamically due to removals. In the second model, each infectious peer is cured and gains immunity such that it does not receive infection again. In this model, there exist only infectious, susceptible and immune peers.

B. Dissemination Algorithms

There are two approaches for epidemic dissemination described as follows.

1) Simple epidemics: In this algorithm, epidemics disseminate from an infectious peer to a subset of its neighbors, defined by the fanout parameter, in each epidemic round. Since there is no mutual exchange of state information, an infectious peer may receive a particular data message multiple times. Hence, this causes redundant message transmission in the network. However, simple epdemics has reduced overhead in comparison to broadcasting/flooding.

2) Anti-entropy (gossip) algorithms: In these algorithms, peers in the network choose one or a group of its neighbors determined by fanout and exchange status information prior to actual data dissemination. This phase is called gossiping. There exist three approaches for information exchange, namely pull, push and hybrid, as particular models of anti-entropy [5]. In anti-entropy algorithms, information carried on each peer is compared prior to information exchange to avoid the pitfall of sending unnecessary information as in simple epidemics. The algorithm causes no overhead but gossiping is a required phase.

III. Related Work

In earlier work [1] using simple epidemics with SIS model, the effect of network topology on dissemination is examined. A critical ratio for detecting if the epidemics will spread to entire network or not is named as epidemic threshold. The average connectivity in the network is denoted by $< k >$, and the connectivity divergence is by $< k^2 >$, the mean and the second moment of the degree distribution, respectively. It has been suggested that an epidemic threshold is $\tau = \frac{1}{<k>}$ for homogenous Erdös-Rényi networks and $\tau = \frac{<k>}{<k^2>}$ for power-law topologies. In [1], a general epidemic threshold of $\tau = \frac{1}{\lambda_{1,A}}$ is suggested for an arbitrary network where $\lambda_{1,A}$ is the largest eigenvalue of the adjacency matrix. It has been shown that infection eventually dies out if $\frac{\phi}{\delta} < \frac{1}{\lambda_{1,A}}$ where ϕ is the infection rate and δ is the cure rate.

In another study again considering SIS [2], strength of the spreading is examined and the role of the topological properties over persistence of the epidemics is emphasized. When n represents the total number of nodes in the network, it has been shown that spreading rapidly takes $O(\log(n))$ rounds when $\frac{\phi}{\delta} < \frac{1}{\lambda_{1,A}}$ and it takes $\Omega(e^n)$ rounds when $\frac{\phi}{\delta} > \frac{1}{\lambda_{1,A}}$.

In [5], SI model and anti-entropy algorithms are considered assuming that each peer has global knowledge of all peers. That is, any other peer in the network can be chosen as a gossip target. Although this assumption is not realistic, it is a crucial simplification for the exact probability calculations performed in [5]. The probability distribution of the number of newly infected peers at each round is derived for push, pull and hybrid algorithms.

IV. Proposed Model

Our model examines epidemic dissemination with pull based anti-entropy algorithm and SI epidemic spreading. The pull algorithm is given below in which spreading data is triggered by susceptible peers (by *pulling* data) when they are picked as gossip destinations by infectious peers. In SI model, the infectious peers are never cured and continue to infect the remaining susceptible peers until the infection is spread over the network as in information diffusion. The analytical model we develop in this section is an extension of earlier work developed for SIS simple epidemic which is used for spreading of viruses in particular and a peer becomes susceptible after a cure [1], [7].

Algorithm 1 Pull Algorithm

Node I is infectious and node S is susceptible. When I picks a neighbor S as the gossip target, infection is triggered:
1. After state exchange via gossip, S requests missing data from I to initiate the pull action.
2. S receives (pulls) the data from I.
3. Upon receiving the data, S becomes infectious.

We derive equations to calculate the infection probability of each peer (node) in consecutive epidemic rounds. The following notation is used:

$p_{i,t}$:probability that node i is infected at time t
$\zeta_{i,t}$:the probability that a node i will not receive infections from its neighbors at time t
n_j :total number of neighbors of a node j, that is,

$$n_j = \sum_{k=1}^{N} A(j,k) \tag{1}$$

where A is the adjacency matrix and N is the total number of nodes.

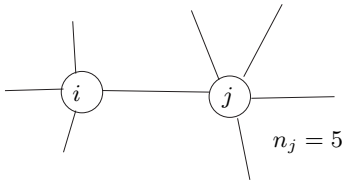

Fig. 1. Node selection

The selection process for a node i by node j in the pull approach is illustrated in Fig. 1 where node j has 5 neighbors and hence i becomes infectious with probability 1/5. Clearly, if there are multiple neighbors of i which are infectious, then the probability of i being selected increases in a given round.

A node i remains susceptible at time t when either one of the following occurs

- neighbor node j is susceptible at time $t-1$, which has probability $1 - p_{j,t-1}$

- neighbor node j is infected at time $t-1$ but chooses a neighbor other than i, which happens with probability $(n_j-1)/n_j$

Since the neighbors act independently in anti-entropy model, we can write the probability that a node i remains susceptible at time t as

$$\begin{aligned}\zeta_{i,t} &= \prod_{j:\text{ neighbor of } i}\left[(1-p_{j,t-1})+\left(p_{j,t-1}\left(\frac{n_j-1}{n_j}\right)\right)\right] \\ &= \prod_{j:\text{ neighbor of } i}\left(1-\frac{p_{j,t-1}}{n_j}\right)\end{aligned}$$

Then, the probability that a node i is susceptible at time t is the product of the probability that it is susceptible at time $t-1$ and the probability that it does not receive infection from its neighbors. That is,

$$1-p_{i,t}=(1-p_{i,t-1})\prod_{j:\text{ neighbor of } i}\left[1-\left(\frac{p_{j,t-1}}{n_j}\right)\right] \quad (2)$$

We show that epidemic will spread to entire network, in other words the system is stable at $\vec{P}=\vec{1}$, irrespective of the size of the initial number of infected node, where $\vec{P}$ is the vector of entries p_i, $i=1,...,n$. It is convenient to work with the probability of being susceptible rather than being infected. Let $q_{i,t}=1-p_{i,t}$. From (2), it is given by

$$q_{i,t}=q_{i,t-1}\prod_{j:\text{ neighbor of } i}\left[\left(1-\frac{1}{n_j}\right)+\left(\frac{q_{j,t-1}}{n_j}\right)\right] .$$

The probability that node i is still susceptible at time t can be represented with the following discrete non-linear dynamical system: $\vec{Q}_t=\vec{f}(\vec{Q}_{t-1})$ with $f=(f_1,\dots\ f_n)$ where

$$f_i(\vec{Q})=q_i\prod_{j:\text{ neighbor of } i}\left[\left(1-\frac{1}{n_j}\right)+\left(\frac{q_j}{n_j}\right)\right]$$

and $\vec{Q}$ is the vector of entries q_i, $i=1,...,n$ after suppressing the time for simplicity. The system's being stable at $\vec{Q}=\vec{0}$ means that the information will certainly diffuse, that is, P_t will converge to $\vec{1}$, starting with any initial number of infectious nodes. Due to [8], pg. 280, the system is stable at $\vec{Q}=\vec{0}$ if the eigenvalues of $\nabla f(\vec{0})$ are less than 1 in absolute value. The gradient matrix is given by the entries $[\nabla f(\vec{Q})]_{ik}=\partial f_i(\vec{Q})/\partial q_k$, $i,k=1,\dots,N$. Taking the partial derivatives, we get

$$\frac{\partial f_i(\vec{Q})}{\partial q_i}=\prod_{j:\text{ neighbor of } i}\left[\left(1-\frac{1}{n_j}\right)+\left(\frac{q_j}{n_j}\right)\right]$$

since $j\neq i$ when j neighbor of i. On the other hand, $\partial f_i(\vec{Q})/\partial q_k=0$ if $k\neq i$ and k is not a neighbor of i since $f_i(\vec{Q})$ doesn't depend on q_k. Finally,

$$\begin{aligned}\frac{\partial f_i(\vec{Q})}{\partial q_k} &= q_i\,\frac{\partial}{\partial q_k}\left[\left(1-\frac{1}{n_k}\right)+\left(\frac{q_k}{n_k}\right)\right] \\ &\quad\cdot\prod_{j:\text{ neighbor of } i,j\neq k}\left[\left(1-\frac{1}{n_j}\right)+\left(\frac{q_j}{n_j}\right)\right] \\ &= \frac{q_i}{n_k}\prod_{j:\text{ neighbor of } i,j\neq k}\left[\left(1-\frac{1}{n_j}\right)+\left(\frac{q_j}{n_j}\right)\right]\end{aligned}$$

as $k\neq i$ when k is a neighbor of i. Therefore,

$$\frac{\partial f_i(\vec{0})}{\partial q_k}=\begin{cases}\displaystyle\prod_{j:\text{ neighbor of } i}\left(1-\frac{1}{n_j}\right) & \text{if}\quad k=i\\ 0 & \text{if}\quad k\neq i\end{cases}$$

In matrix notation, we find

$$\nabla f(\vec{0})=\text{diag}(\lambda_1,\dots,\lambda_N)$$

with

$$\lambda_i=\prod_{j:\text{ neighbor of } i}(1-1/n_j)\qquad i=1,\dots,N.$$

Clearly, λ_i are simply eigenvalues of $\nabla f(\vec{0})$ and $0\leq\lambda_i<1$. Therefore, the information will certainly diffuse as expected.

The analysis above does not only confirm the applicability of the discrete model (2) for epidemic diffusion, but also provides the tools for evaluating the rate of dissemination in connection with the adjacency matrix. Scrutinizing the stability proof of [8] which states that there exists a constant $\mu<1$ such that

$$\|\vec{Q}_t\|\leq\mu^t\,\|\vec{Q}_0\| \quad (3)$$

we see that μ can be chosen as a perturbation $|\lambda|+\epsilon$ of the maximum eigenvalue λ in magnitude of $\nabla f(\vec{0})$ where $\epsilon>0$ can be chosen arbitrarily small. The largest eigenvalue would be binding in the worst case, especially for large t. Therefore, Equation (3) reflects that the dissemination occurs exponentially with a rate depending in general on all the eigenvalues $\lambda_1,\dots,\lambda_N$ which are found above in terms of the row sums (1) of the adjacency matrix. Since (1) corresponds to the number of degrees of each node j, we explore the effect of the degree distribution as well as the eigenvalues on the diffusion rate for different random topologies next.

V. Numerical Results

We consider power-law and Erdös-Rényi graphs as overlay topologies. Power law graphs have attracted great interest since the Internet topology exhibits a power law degree distribution. A power law graph is one where the number of nodes with degree k is proportional to $k^{-\beta}$ for some $\beta>1$. For the mean degree to be finite, we need $\beta>2$. On the other hand, Erdös-Rényi graph is of interest as a bench-mark random graph. Erdös-Rényi is characterized by parameters n and p where n is the number of nodes, and there exists an edge between each pair of nodes with probability p independently from the other edges. It follows that the average degree is $(n-1)p$ [2].

We evaluate epidemic spreading in various power-law graphs using Barabási power-law graph generator [9]. The nodes have an average degree which is twice of a free parameter in the generator. The algorithm creates networks with a distribution following $k^{-2.9\pm0.1}$. For Erdös-Rényi graphs, we vary the parameter p to obtain different mean degrees. The network size is 1024 and we evaluate 10 graphs of each topology by varying the mean degrees. The expected number of infected nodes is found by adding the entries of the vector P_t and we report the percentage of infected nodes in our numerical evaluations. The mean degree and the eigenvalues of the gradient matrix have been investigated with respect to the rate of diffusion. We examine the percentage of infected nodes at 15th and 20th rounds of dissemination. At the 15th round, infection disseminates significantly on the graph and at the 20th round the dissemination is almost complete.

As observed in Fig.2, the diffusion rate increases quickly with the mean degree up to a certain threshold, in this case 10, then only slightly for larger degrees. Erdös-Rényi graphs show faster dissemination when compared with power-law graphs with the same mean values. Since the rate differs for the same mean values, we conclude that mean degree is not a discriminating graph invariant across different topologies.

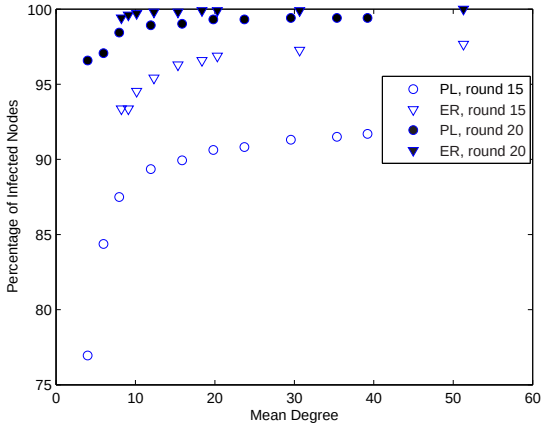

Fig. 2. Impact of mean degree on diffusion

We have observed that the mean of eigenvalues of the gradient matrix discriminates the groups of different topologies at both 15th and 20th rounds. Erdös-Rényi graphs all have a mean about 0.37 while power-law graphs have mean eigenvalue of 0.43 and larger as shown in Fig.3.

We report the standard deviation of the eigenvalues in Fig.4 which distinguishes clearly both between groups and within a specific group. Erdös-Rényi graphs all have smaller deviation of eigenvalues compared to power-law. In general, dissemination rate is inversely proportional to mean and standard deviation of the eigenvalues. We have also investigated the standard deviation of the degree distribution. We conclude that Erdös-Rényi graphs show faster dissemination when compared with power-law graphs

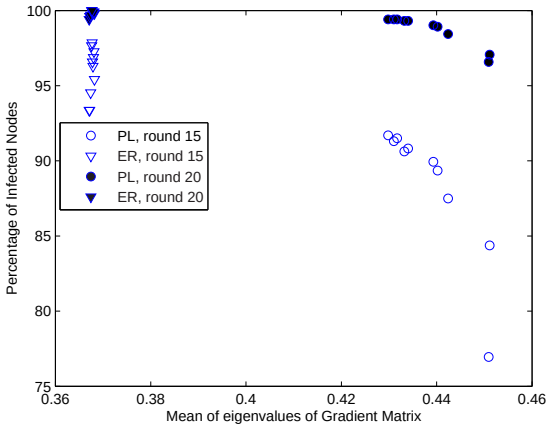

Fig. 3. Impact of mean of eigenvalues of gradient matrix

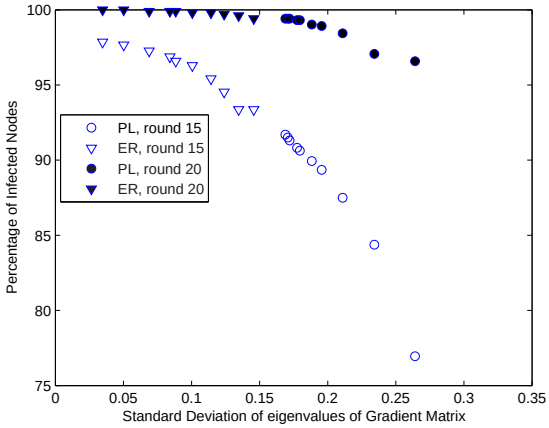

Fig. 4. Impact of standard deviation of eigenvalues of gradient matrix

since they have smaller standard deviation for both eigenvalue and degree distributions.

The maximum eigenvalue of the gradient matrix depicted in Fig.5 shows a similar behavior to mean degree given in Fig.2. Therefore, maximum eigenvalue alone is not a discriminating factor for different random graphs. Indeed, Erdös-Rényi graphs show faster dissemination when compared with power-law graphs with the same maximum eigenvalues.

VI. Conclusion

We have derived an analytical model for pull type anti-entropy approach for SI epidemic information dissemination. We have assumed neighborhood knowledge among peers and information exchange based on proximity. Our model explicitly involves overlay topology through the inclusion of its adjacency matrix. The rate of dissemination is found to be related to the adjacency matrix in a nonlinear way. However, we can explicitly compute the gradient matrix of

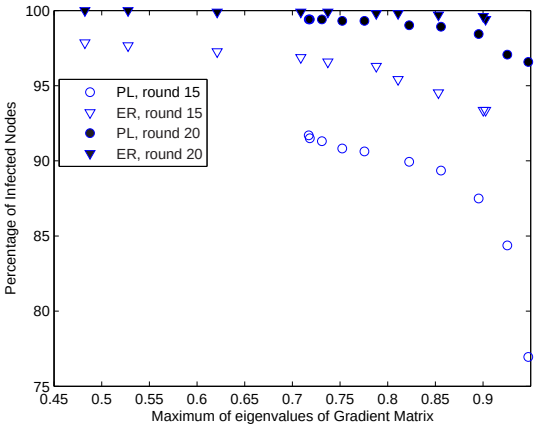

Fig. 5. Impact of maximum of eigenvalues of gradient matrix

the function that governs the dynamics of diffusion. In our numerical evaluations, we have investigated the topological properties such as degree distribution and eigenvalues of the gradient matrix over Erdös-Rényi and power-law random graphs. Rather than the maximum eigenvalue, the mean and the standard deviation of all eigenvalues are found to be effective in predicting the rate of diffusion.

ACKNOWLEDGMENT

This work is supported in part by TUBITAK (The Scientific and Technical Research Council of Turkey) under CAREER Award Grant 104E064.

REFERENCES

[1] Y. Wang, D. Chakrabarti, C. Wang and C. Faloutsos, *Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint*, Proc. IEEE SRDS, 2003.
[2] A. Ganesh, L. Massouli and D. Towsley, *The Effect of Network Topology on the Spread of Epidemics*, Proc. of IEEE INFOCOM, 2005.
[3] B. Cohen, *Incentives build robustness in BitTorrent*, P2P Economics Workshop, Berkeley, CA, 2003.
[4] A. Alagöz, Ö. Özkasap and M. Çağlar, *Principles and Performance Analysis of SeCond: A System for Epidemic Peer-to-Peer Content Distribution*, submitted.
[5] Ö. Özkasap, E. Ş. Yazıcı, S. Küçükçifçi and M. Çağlar, *Exact Performance Measures for Peer-to-Peer Epidemic Information Diffusion*, LNCS 4263, Proc. of ISCIS, 2006.
[6] M. Draief, *Spread of Epidemics and Rumours in Networks*, UK Social Network Conference, 2007.
[7] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec and C. Faloutsos, *Epidemic Thresholds in Real Networks*, ACM Transactions on Information and System Security, 10: 1-26, 2008.
[8] M. W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, 1974, Academic Press.
[9] http://www.cs.ucr.edu/ ddreier/barabasi.html