Network protection against worms and cascading failures using modularity partitioning

Jasmina Omić
Network Architectures and Services
Delft University of Technology
Delft, Netherlands
Email: J.Omic@tudelft.nl

Javier Martín-Hernández
Network Architectures and Services
Delft University of Technology
Delft, Netherlands
Email: J.MartinHernandez@tudelft.nl

Piet Van Mieghem
Network Architectures and Services
Delft University of Technology
Delft, Netherlands
Email: P.F.A.VanMieghem@tudelft.nl

***Abstract*—Communication networks are prone to virus and worms spreading and cascading failures. Recently, a number of social networking worms have spread over public Web sites. Another example is error propagation in routing tables, such as in BGP tables. The immunization and error curing applied to these scenarios are not fast enough. There have been studies on the effect of isolating and curing network elements, however, the proposed strategies are limited to node removals.**

This paper proposes a link isolation strategy based on the quarantining of susceptible clusters in the network. This strategy aims to maximize the epidemic control while minimizing the impact on the clusters performance. We empirically study the influence of clustering on robustness against epidemics in several real-world and artificial networks. Our results show an average curing rate improvement above 50% for the studied real-world networks under analysis.

I. Introduction

Epidemics on networks, from worm epidemics in computer networks to information spread in P2P and ad-hoc networks [13], [28] have recently attracted a lot of attention.

After the scanning worms, a new challenge for network security is posed by the strain of worms that use social networking Websites to spread. Web applications for exchange of information and data introduced new vectors of spread. Many social network worms use AJAX[1] scripts like Samy [19], Yamanner [7] and Mikeyy [20]. Worm spreading usually involves user interaction in order to download worm payload on the local machine as Koobface [4], but recently Web clients are infected simply by visiting a Web page; no user interaction is necessary [19]. The infection risk increases since social networks are not restricted only to Facebook and Twiter, but are becoming embedded in other not strictly social websites like Digg and Youtube. Additionally, social networks have power law network structure which makes them prone to epidemic spreading [28], [3], [17] and [26].

The epidemic algorithms for information dissemination in unreliable distributed networks such as P2P and ad-hoc networks show similar epidemic dynamics on networks [13], [5]. Finally, the propagation of faults and failures can be modeled as an epidemic. Coffman *et al.* [18] models cascading BGP failures on a fully connected topology. We concentrate on protection against worms and error propagation in communication networks.

The protection of important networks in the above mentioned cases is in practice not fast enough, and the infection easily reaches all the segments of the network. This paper proposes and analyzes a fast method to stop or reduce epidemic spreading on networks. When an epidemic is detected, a network cut is performed by removing links leading to several disconnected clusters of nodes. This clustering allows limited intercommunity communication between nodes to continue, while possibly quarantining the rest of the network. Many real-world networks from on-line social networks to airline transport networks and Internet ASes network typically show a strong community structure [15], [21]. Depending on the speed of the epidemic reaction, it is possible to totally prevent any risk of infection for a number of disconnected clusters. Even with very delayed reaction, the amount of protection, that has to be applied in the network in order to stop the spreading, can be reduced. Thus, clustering can be used in addition to other protection methods.

The removal of links as protection against epidemics was proposed in mathematical epidemiology. The Equal Graph Partitioning (EGP) method uses immunization to remove specific nodes that cut the graph into clusters [6]. However, the immunization takes time, while individual nodes can stop communicating with other nodes immediately after receiving the news about the epidemic. Several authors have studied the reduction of disease spreading using air line restrictions. Goedecke *et al.* [16] and Epstein *et al.* [22] used the Susceptible Exposed Infected Recovered ($SIER$) model and dynamic time travel restrictions. Marcelino *et al.* [21] used the Susceptible Infected (SI) model together with edge betweenness and the Jaccard coefficient to increase the spreading time [21] by 81% by removing 25% of the links. Due to the multicommunity structure of the network with most connected nodes not being the most central, the optimal strategy for flight cancellation is not the removal of nodes (cities), but the removal of intercommunity flights, which introduced an increase in spreading time [21]. We are interested in specific link removal such that intracommunty communication is preserved. We are not interested in optimizing of clustering algorithm, but instead in the general improvement of protection that is possible by using a well-

[1]Asynchronous JavaScript and XML

defined clustering algorithm.

Several algorithms have been proposed to find network communities. Modularity maximization is the most popular method. The modularity Q is a quantitative criterion to evaluate how good a graph partition is [25]. It maximizes links within communities, while minimizing the links between them. Modularity maximization is an NP problem, given the exponential number of possible partitions. In this paper, we use a greedy heuristic proposed by Clauset *et al.* [8] to find an optimal modularity clustering.

In order to quantify the improvements of the network clustering in terms of epidemics, we use the epidemic threshold concept and the N-intertwined Susceptible Infected Susceptible (SIS) epidemic model [24] on a large set of networks. In a SIS epidemic model, the epidemic can be stopped, provided the network protection functionalities against the virus perform faster than the reproduction of the virus. The epidemic thus exhibits threshold behavior.

In section II, we explain the protection algorithm and describe the networks that we examine. The epidemic theory used to estimate the protection is explained in section III. Results are presented in section IV, with comparison of random link removal and the modularity algorithm in section IV-C.

II. Quarantine model and networks

The protection method of dividing the network into clusters by removing links will be referred to as *clustering* or *quarantining*. The moment when a network is quarantined determines how many nodes are completely protected, since the virus is not able to infect nodes outside its cluster. In the first case, if we are able to quarantine a network into clusters faster than the virus is spreading, only a single cluster will contain infected nodes. On the other hand, if the virus infects all the clusters before a quarantine takes place there are still benefits, which are discussed in more details in section IV-B. Usually, the effective speed of clustering the network will be somewhere in between.

We discuss the two boundary cases separately. In the first case we determine the size of the clusters, which provides an estimate of how many nodes will never get infected. The size of the clusters also affects the performance of the network. Larger clusters mean that a larger part of the network can continue exchanging information. Second, we show that the epidemic threshold that divides non-infected from infected networks favorably increases in networks that display clustering features.

If the infection is spreading very fast and all the clusters get infected, the number of infected nodes in the metastable state is reduced. We discuss the improvement with the respect to the number of removed links.

To illustrate the influence of clustering on epidemic spreading, we use several real-world networks. First, the Internet AS level topology obtained by Route View in 2006 and posted by the University of Oregon is used to demonstrate the effect of clustering on the virus spread in large infrastructural networks.

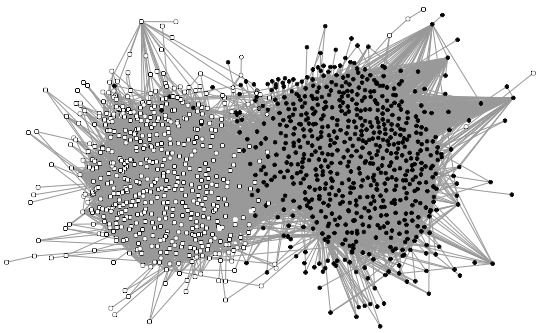

Fig. 1. Network of weblogs on US politics clustered network using modularity maximization. Nodes belonging to different clusters are differently colored.

Further, we used an example of a social network between weblogs on US politics recorded in 2005 by Adamic and Glance [1]. The political blog network is shown in Fig. 1, with nodes belonging to different clusters colored in different colors. Finally, we examine an on-line social network of friends from www.digg.com, collected by the NAS at Delft University of Technology.

In disease modeling, transport networks are frequently used. To illustrate the influence of traveling patterns on virus spread, we investigate the direct airport-to-airport American traffic network maintained by the U.S. Bureau of Transportation Statistics and the European direct airport-to-airport traffic network obtained from the European commission for statistics Eurostat. The number of links and nodes of different networks are given in Table II.

In order to extend our understanding of the effects of clustering on the network robustness against virus spread, we include several artificial networks with $N = 1,000$ nodes.

We consider three Erdös-Rényi (ER) random graphs with a different number of links. Each node in the ER random graph is connected to every other node with probability p. The probability p determines the number of links in the network [12]. We model power law networks using the Barabási-Albert model (BA) of preferential attachment for different number of links [2]. Finally, we use an artificial model of clustered networks [27]. The network is constructed in a similar manner as the ER random graph with two probabilities of link existence, one for inter-community connections and the other for intra-community connections. We have generated several different networks with $N = 1,000$, two clusters and different modularity. Further, we have considered networks with $4, 6, 8, 10$ clusters. We choose to generate a greater number of networks with two clusters because most of the real-world networks consist of mainly two big clusters.

We additionally consider the square lattice, line, ring and tree topologies.

The networks are not weighted; however, the N-intertwined model is extendable to a heterogeneous setting [23].

III. N-INTERTWINED MODEL EPIDEMIC THRESHOLD

To model epidemic spread, we use the N-intertwined SIS model, which was introduced and discussed in [24]. A SIS model is one of the standard epidemic models: a node is susceptible to infection (S), then it becomes infected (I) and, after curing, it is susceptible to infection (S) again.

In order to quantify reduction of the number of infected nodes gained by clustering in the case of slow separation of the network, we use results of the N-intertwined model. A network is modeled as a connected, bidirectional graph $G(N, L)$.

By separately observing each node, the infection spread is modeled in a bidirectional network specified by a symmetric adjacency matrix A. A node i at time t can be in one of the two states: *infected*, with probability $v_i(t) = \Pr[X_i = 1]$ or *susceptible*, with probability $1 - v_i(t)$. The sum of the probabilities of being infected and susceptible are equal to 1 because a node can only be in one of these two states. The state of a node i is specified by a Bernoulli random variable $X_i \in \{0, 1\}$: $X_i = 0$ for a susceptible node and $X_i = 1$ for an infected node. We assume that the protection process per node i is a Poisson process with rate δ, and that the infection per link is a Poisson process with rate β which is imminent for all nodes and thus constant in the network. For a node i, we can formulate the following differential equation

$$\frac{dv_i(t)}{dt} = \beta(1 - v_i(t))\sum_{j=1}^{N} a_{ij}v_j(t) - \delta v_i(t)$$

where a_{ij} is the element of the adjacency matrix A and it is equal to 1 if the nodes i and j are connected, otherwise it is 0. A node is not considered connected to itself, i.e. $a_{ii} = 0$. The probability of a node being infected depends on the probability that it is not infected $(1-v_i(t))$ multiplied with the probability that a neighbor j is infected $a_{ij}v_j(t)$ and that it tries to infect the node i with the rate β. Detailed derivations are given in [24] and [23].

In the steady-state, where it holds that $\frac{dv_i(t)}{dt} = 0$, and $\lim_{t\to\infty} v_i(t) = v_{i\infty}$ for each node $1 \le i \le N$, we have that

$$v_{i\infty} = \frac{\beta \sum_{j=1}^{N} a_{ij}v_{j\infty}}{\beta \sum_{j=1}^{N} a_{ij}v_{j\infty} + \delta} \tag{1}$$

This system of equations has $2N$ solutions with one positive solution and one solution equal to 0 [24]. The positive solution gives the probability that nodes are in the infected state during the steady-state of the model. The model gives a good approximation of the real epidemic process and the metastable state [24] for a wide range of effective spreading rates $\tau = \frac{\beta}{\delta}$. Thus, we will refer to the metastable state as a steady-state.

The fraction of infected nodes at any given time t can be calculated as a sum of probabilities that the nodes are infected

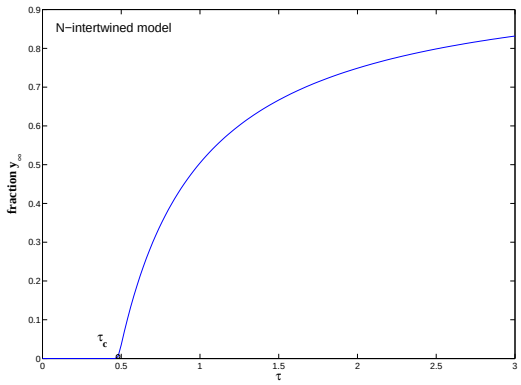

Fig. 2. Fraction of infected nodes as a function of the effective infection rate τ. The epidemic threshold is denoted by τ_c.

$y(t) = \frac{1}{N}\sum_{j=1}^{N} v_j(t)$ and in the steady-state $y_\infty = \frac{1}{N}\sum_{j=1}^{N} v_{j\infty}$.

For a fixed curing rate and spreading rate, the fraction of infected nodes as a function of the effective spreading rate τ is given in Fig. 2. The model as well as the real epidemic process have a threshold value at τ_c. The threshold can be defined as follows: for effective spreading rates (rate of spread divided by rate of protection) below some critical value the virus in the network with N nodes dies out before a large population is infected with a mean epidemic lifetime of the order of $O(\log N)$. For effective spreading rates above the critical value τ_c, the epidemic persists and the number of infected nodes is large, with a mean epidemic lifetime [14] of the order of $O(e^{N^\alpha})$ for a SIS model. The state above the epidemic threshold is referred to as the *metastable state*. In the metastable state, some constant mean portion of nodes is infected [24].

The epidemic threshold is equal to $\tau_c = \frac{1}{\lambda_{\max}(A)}$, where $\lambda_{\max}(A)$ is the largest eigenvalue of the matrix A [14], [29] and similar results exist for Susceptible Infected Removed SIR model [11], [10]. We denote $\lambda_{\max}(A)$ with $\lambda_{\max G}$.

If $\tau < \tau_c$, the infection will eventually be cured, and for $\tau > \tau_c$ the infection persists with the average number of infected nodes equal to y_∞.

For example, the largest eigenvalue of a line graph is $\lambda_{\max G} \simeq 2$, while that of a star topology is $\lambda_{\max G} = \sqrt{N-1}$. These two graphs are interesting examples, because both have the same number of links $L = N - 1$. Thus, the spreading in a star topology is significantly higher than in a line topology with the same number of nodes and links.

Figure 2 shows the threshold behavior for the steady-state of an infected network.

IV. RESULTS

In this section, we examine the case of instant clustering where a network is clustered faster than the worm is spreading, resulting in a single infected cluster. Further, we consider the case where all the clusters are infected before the quarantine

process clustered the network. Finally, we compare the quarantined networks with networks where the same number of links has been randomly removed.

A. Early clustering

Defending the network and performing quarantines provides important advantages. First of all, if a network is cut on time and the infection is limited to one cluster, only a percentage of nodes will eventually be exposed to infection. Second, from the interlacing theorem of graph theory [9], the largest eigenvalue of a subgraph is always smaller than that of the graph. Thus, the thresholds $\tau_c = 1/\lambda_{max}$ will always increase for any subgraph, making the subgraphs more robust against epidemic spreading. The case that all the clusters are initially infected is discussed in section IV-B. Finally, the lifetime of the metastable state depends on the number of nodes [14] as $\Omega(e^{N^\alpha})$, for $\alpha > 0$. The number of removed links using the modularity algorithm ranges from 7% to 58% of the links. The values for different networks are given in Table II.

One of the improvements introduced by clustering is a reduction of the largest eigenvalue $\lambda_{\max}$ of the smaller clusters with respect to the original graph. This increases the threshold τ_c, the border between infected and non-infected networks. The ratio between the largest eigenvalue of a cluster and the largest eigenvalue of the whole network versus the modularity Q for several networks is shown in Fig. 3 and 4.

The behavior of $\lambda_{\max Cluster}$ for the different network is diverse. For networks with high modularity, such as the lattice and tree topologies, the improvement, a lowering of $\frac{\lambda_{\max Cluster}}{\lambda_{\max G}}$ is not so significant. For the same type of networks e.g. BA or ER with different number of links, a reduced modularity results in a reduced $\lambda_{\max}$, which is an improvement. For both cases, the modularity is reduced by generating topologies with a larger number of links (by respectively increasing the parameter m in the BA model and the parameter p in the ER model). In addition, the difference between the two largest eigenvalues of different clusters is greater for BA than for ER. The effect can be caused by the homogeneity of the degree distribution of clusters in the ER case, while BA shows a significantly heterogeneous cluster degree distribution.

The threshold $\tau_{c,cluster} = \frac{1}{\lambda_{\max Cluster}}$) increases as a function of the number of links removed between a cluster and the rest of the network, as shown in Fig. 5 and 6. In order to preserve as much network communication as possible upon link removal, a small number of links should be removed during the quarantine. On the other hand, τ_c is inversely proportional to $\lambda_{\max Cluster}$. Hence the networks with best performance show clusters with both low $\lambda_{\max Cluster}$ and low L_{out}, close to the point $(0, 0)$ in the figures. Real-world networks such as the airline networks and AS network perform well, while artificial networks perform much better the smaller the number of clusters in the graph is.

For individual graphs, the dependency of threshold improvement versus the number of links removed is close to linear, which is indicated by change in lower bound on largest eigenvalue $\lambda_{\max} \geq \frac{2L}{N}$. Sparse ER graphs are clustered easily,

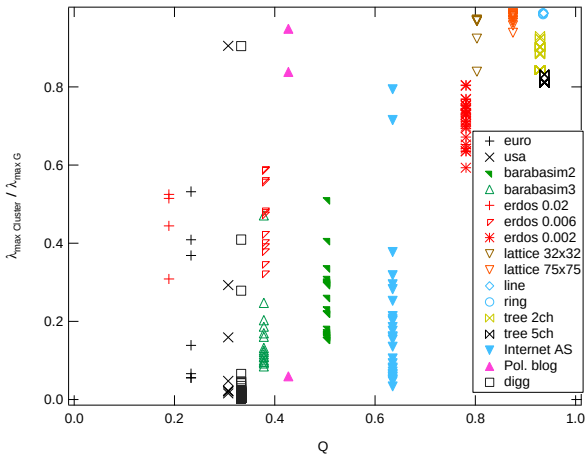

Fig. 3. Relative largest eigenvalue of the each cluster $\lambda_{max\,Cluster}/\lambda_{max\,G}$ as a function of the modularity Q for real-world networks and real-world models.

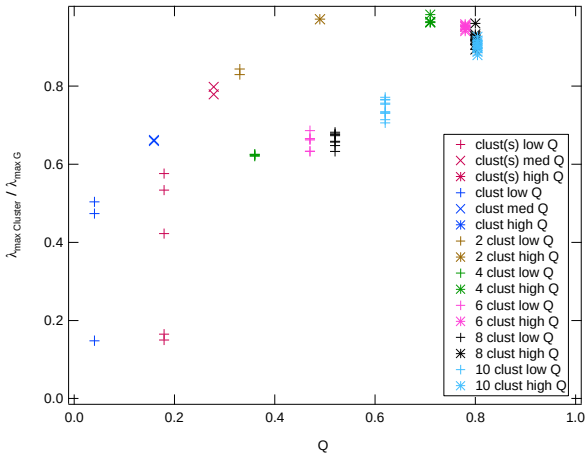

Fig. 4. Relative largest eigenvalue of the each cluster $\lambda_{max\,Cluster}/\lambda_{max\,G}$ as a function of the modularity Q for cluster network models.

with a small number of removed links, but show no significant improvement in τ_c. The artificial, clustered graph with low modularity shows the worst performance in the number of removed links, as in Figure 4.

The size of the clusters after cutting is an important variable for the performance of the network. Large clusters will allow for node communication after a quarantine. But on the other hand smaller clusters will be more robust to virus spread. The size of the clusters is decided by the modularity algorithm.

Another parameter to consider is the size of the largest cluster after the quarantine. The distribution of the fraction of cluster sizes $\frac{N_c}{N}$ is shown in Fig. 7. In the case of early clustering, the network is cut into clusters before the virus can reach any other cluster except for the one it starts to spread in. The worst case scenario is when the virus starts to spread in the largest cluster. Most of the networks have one cluster that contains half of the nodes. In the case of the European air network, the three clusters pop up, thus leaving more than

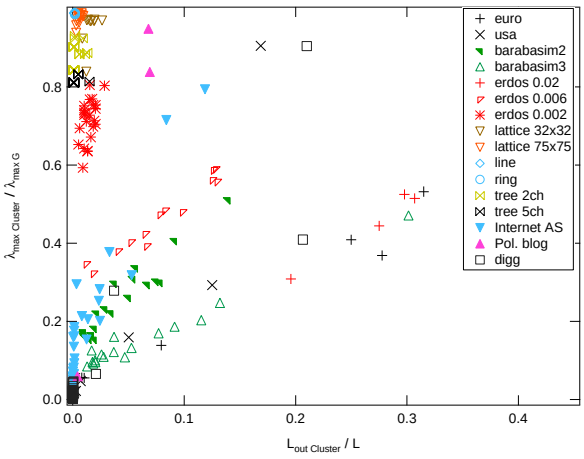

Fig. 5. Relative largest eigenvalue of the cluster $\lambda_{max\,Cluster}/\lambda_{max\,G}$ as a function of the relative number of links leaving the cluster $L_{out\,Cluster}/L$ real-world networks and real-world models.

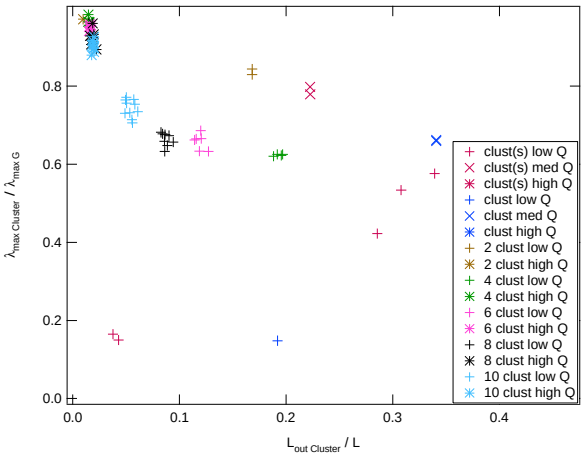

Fig. 6. Relative largest eigenvalue of the cluster $\lambda_{max\,Cluster}/\lambda_{max\,G}$ as a function of the relative number of links leaving the cluster $L_{out\,Cluster}/L$ for cluster network models.

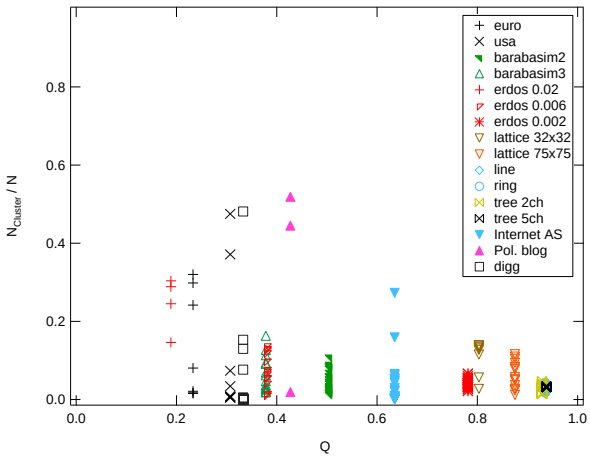

Fig. 7. Relative number of nodes in the cluster $N_{Cluster}/N$ as a function of the modularity Q real-world networks and real-world models.

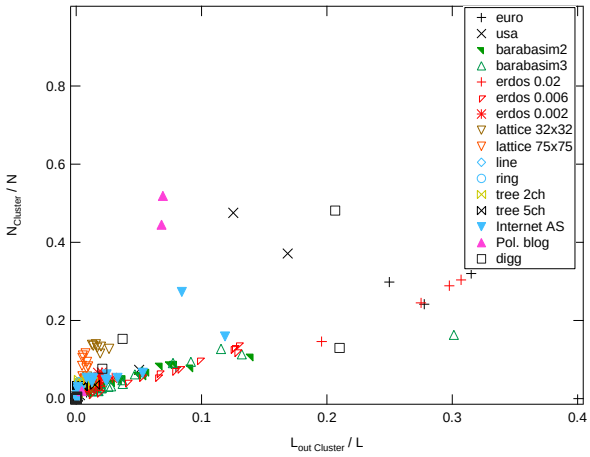

Fig. 8. Relative number of nodes in the cluster $N_{Cluster}/N$ as a function of the relative number of links leaving the cluster $L_{out\,Cluster}/L$ for real-world networks and real-world models.

two thirds of network protected. A BA graph has many small clusters of the size one fifth of network, which leaves four fifths of network protected, as shown in Fig.7.

The Digg network has one large cluster which covers half of the network and many significantly smaller ones. The USA air network and the political blog network have 2 large clusters, while the European air network has 3 large clusters and several small ones. The Internet AS topology is more differentiated. There are 8 clusters with $1,000 - 1,500$ nodes and two larger ones with $3,000$ and $6,000$. Artificial networks show different behavior. The ER and BA network have a lot of smaller clusters comparable in size. In Fig. 8, the number of nodes in the cluster is given as a function of the number of removed links between the cluster and the rest of the network. The air network of USA airports has the largest cluster with the smallest number of deleted links, while the European air network has 3 clusters.

In Fig. 9 and 10, for the same network, larger clusters tend to have a larger $\lambda_{\max Cluster}$ than the smaller clusters. This is, however, not true for any graph: compare the path graph of any size with the complete graph of any smaller size.

B. Delayed clustering

We examine the number of infected nodes using the N-intertwined model. In order to clean the infected network, it is necessary to apply a protection/cleaning rate δ such that the effective spreading rate $\tau = \frac{\beta}{\delta}$ is below the threshold $\frac{1}{\lambda_{max}}$. If the network is completely infected and then clustered, the amount of cleaning is reduced because $\lambda_{\max Cluster} \leq \lambda_{\max G}$, therefore $\tau_c(G) \leq \tau_c(Cluster)$. Thus, if the network is clustered, it will be easier to clean the network from infection.

Fig. 11 presents the percentage of infected nodes as a function of the effective spreading rate τ for different clusters in the artificial, cluster network with low modularity Q.

We calculate the fraction of infected nodes in the clustered network y_{clust} for the effective spreading rate τ for which the number of infected nodes in the original network $y_{tot,50\%}$,

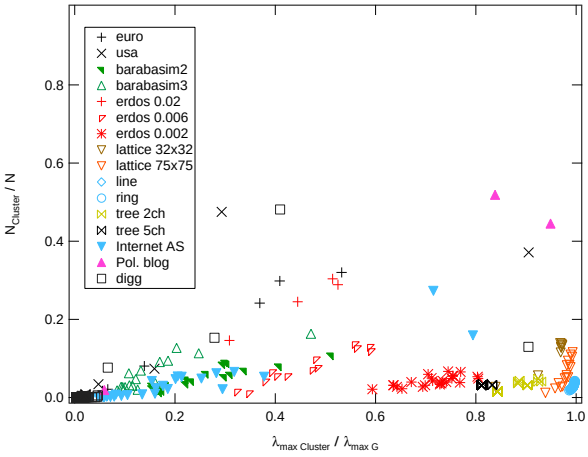

Fig. 9. Relative number of nodes in the cluster $N_{Cluster}/N$ as a function of the relative largest eigenvalue of the cluster $\lambda_{max\,Cluster}/\lambda_{max\,G}$ for real-world networks and real-world models.

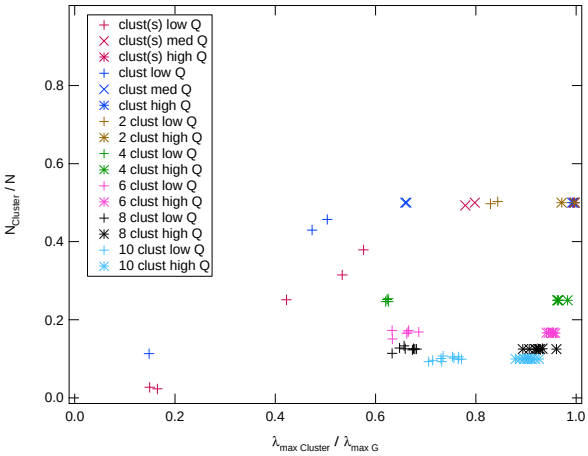

Fig. 10. Relative number of nodes in the cluster $N_{Cluster}/N$ as a function of the relative largest eigenvalue of the cluster $\lambda_{max\,Cluster}/\lambda_{max\,G}$ for cluster network models.

$y_{tot,80\%}$ reaches 50% and 80%. Then, we calculate the difference between the original value and the improved one:

$$i_{50\%} = y_{tot,50\%} - y_{clust}, i_{80\%} = y_{tot,80\%} - y_{clust}$$

We calculate the fraction of infected nodes for several networks. Larger networks as the Internet AS and the Digg network are computationally more demanding and are left out of the analysis. In Fig. 12, the upper bound on the reduction of infected nodes exhibits the tendency to decrease with the modularity of the graph. The improvement is different when there are 50% and 80% of infected nodes in the original network. Air travel networks and ER networks with small average degree do not show significant difference between improvements and have generaly small improvements.

The number of infected nodes decreases with the increase of the number of removed links in the hole network, shown in Fig. 13. This is not surprising because the power of spreading in a network decreases with links removal. Real-

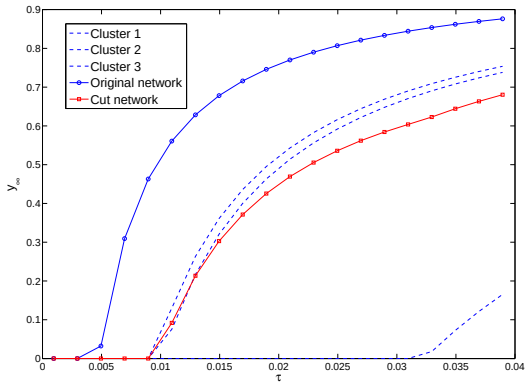

Fig. 11. Percentage of infected nodes y_∞ as a function of the effective spreading rate τ for original network and clustered network.

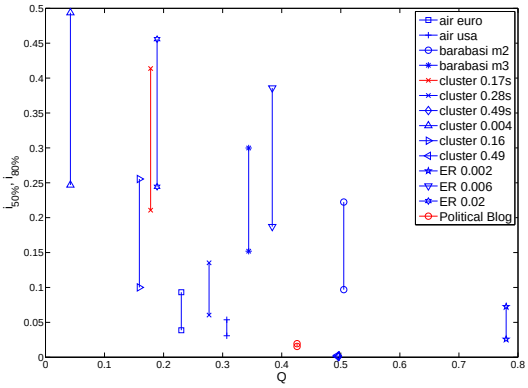

Fig. 12. The difference between the number of infected nodes in the original network and the clustered network as a function of the modularity in the case when 50% and 80% of nodes are infected in original network.

world networks do not show a significant reduction in number of infected nodes.

C. Random removal of nodes

In this section, we compare the threshold τ_c between quarantined networks with networks where the same number of links has been randomly removed. We give the largest eigenvalue of the original graph $\lambda_{\max G}$, the size of the giant connected component $\frac{N_{rand.big.comp}}{N_G}$, its largest eigenvalue $\lambda_{\max rand}$, the size $\frac{N_{big.clust}}{N_G}$ and the largest eigenvalue $\lambda_{\max l.Clust}$ of the largest cluster in the clustered network in Table I. Links are removed at random and the average over many simulations of the largest eigenvalue of the largest connected component is calculated together with the variance of the largest eigenvalue.

The results are presented in Table I. A large part of the network remains connected and can transmit infection, which is an expected result of random link removal. Between 80% and 90% of the network can be affected compared with at most 50% in case of clustering. Further, the largest eigenvalue of the

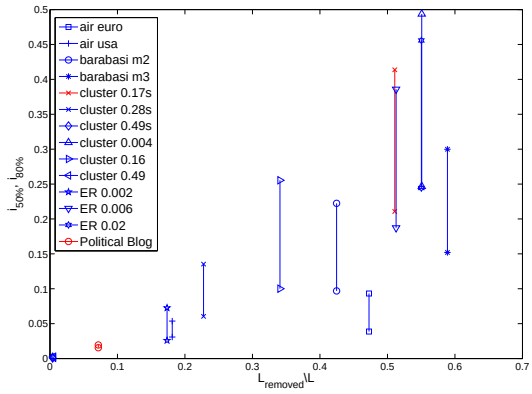

Fig. 13. The difference between the number of infected nodes in the original network and the clustered network as a function of the relative number of removed links in the case when 50% and 80% of nodes are infected in the original network.

Network	N	L_{tot}	$L_{removed}\%$
Euro	$1,247$	$14,952$	47.27%
USA	$2,179$	$31,326$	18.11%
BA 2m	1000	$1,971$	42.46%
BA 3m	1000	$2,673$	58.88%
ER 0.002	808	980	17.34%
ER 0.006	1000	$3,054$	51.27%
ER 0.02	1000	$9,938$	55.02%
AS '06	$22,963$	$48,436$	20.62%
Pol. Blog	$1,222$	$19,021$	7.16%
Digg	$281,471$	$4,354,174$	25.02%

TABLE II
NETWORK COST, THE NUMBER OF REMOVED LINKS.

largest cluster is still smaller than that of the large component in the case of random link removal.

In USA airlines network, ER graphs with $p = 0.002$ and $p = 0.006$ some smaller cluster have larger $\lambda_{\max Cluster}$. In ER graphs and the political blog, two or more components similar in size have almost the same largest eigenvalue. In the case of the political blog the advantage of clustering over random link removal lies in the fact that the other half of the nodes will not get infected if the clustering is performed before the virus has spread. In the case of the AS Internet topology, the smaller cluster of $N = 3,600$ nodes also has a larger $\lambda_{\max Cluster}$ than the largest cluster of $6,200$ nodes. The Digg network also has smaller cluster of $N = 36,491$ nodes with the largest eigenvalue $\lambda_{\max Cluster} = 701.61$, while all the rest of the network has significantly smaller largest eigenvalue. In the case of cluster 28s and 49s, two disconnected components have the same largest eigenvalue, which is the same as for random removal.

The variance of the largest eigenvalue for different simulations of random link removal is less than 0.2 in all cases.

D. Discussion of results

When dividing the network into clusters, a virus can be stopped and annihilated faster. However, protection comes with a cost. Shutting down links from the network reduces the communication and reachability of nodes in the network. Assuming that the graph is disconnected only temporally, we calculate the price of quarantine as the number of links that are removed from the graph as a result of a modularity clustering.

The number of removed links varies from $0,4\%$ to 60%. Most of the considered networks have around 50% of removed links which is significant. In networks where a small number of links is removed, no significant improvement in the largest eigenvalue and the number of infected nodes is found in the steady-state.

Although the modularity maximization algorithm is popular [25], it has not passed a rigorous theoretical examination. The question is also how good its resulting clustering is. We have not examined other algorithms that may perform differently, because we have concentrated on keeping the communities intact.

The largest eigenvalue improvement using the modularity algorithm is comparable with random links removal for several networks; however, in this case the worm can spread to 90% of the network.

V. CONCLUSION

This paper combines the diverse concepts of network clustering, graph spectra and epidemic spread in order to improve the protection against the spread of malware. We have found that real-world networks tend to show a better epidemic threshold τ_c after clustering than artificially generated graphs.

For all the networks under study, the curing rate can improve between 29% an 83% for the largest connected component with respect to the original graph. This wide range of values demonstrates the effect of the network topology on the virus spread. Regarding the network clustering features, an easily clustered graph does not guarantee a slower epidemic threshold, but the way the links intertwine between inter- and intra-communities are key.

Overall, network protection against cascading failures can be improved for any kind of graph. However, the number of removed links is, in practice, unacceptably high. The advantages of early quarantine are shadowed by the fact that up to half of the links must be shut down for the quarantine to take effect.

The real-world networks have typically two or three big clusters and several smaller ones, while BA and ER graphs have several smaller ones comparable in size. BA and ER graphs are assumed to model the real-world complex networks. However, in respect of the size of the clusters, BA and ER fail to match real-world networks.

Additional to the epidemic spread analysis, this diversity in results appears valuable to create a general classification of types of networks. The degree distribution of the graph has been so far widely used for this purpose. For instance, a network classification could be generated by taking the largest eigenvalue of the adjacency matrix of clusters $\lambda_{\max Cluster}$ vs. links that are removed L_{out} as an input.

Network	$\lambda_{\max} G$	$\frac{N_{rand.big.comp}}{N_G}\%$	$\lambda_{\max} rand$	$\frac{N_{big.clust}}{N_G}\%$	$\lambda_{\max} l.Clust$
Euro	80.92	83.23	53.48	31.99	43.07
USA	144.61	96.51	118.67	47.54	42.36
BA 2m	16.09	85.50	12.01	10.08	8.22
BA 3m	28.11	88.40	20.41	16.30	13.24
Cluster 0.17s	22.88	100	11.77	37.9	13.17
Cluster 0.28s	23.51	100	18.41	50.0	18.77
Cluster 0.49s	25.32	100	25.23	50.00	25.26
ER 0.002	3.59	83.41	3.29	6.68	2.67
ER 0.006	7.23	93.2	4.29	13.7	4.03
ER 0.02	20.93	100	10.05	30.4	10.77
AS '06	71.61	90.59	58.49	27.27	51.22
Pol. Blog	74.08	99.01	69.88	51.88	62.11
Digg	775.33	92.7	582.11	48, 13	317.32

TABLE I
COMPARISON OF THE RANDOM LINKS REMOVAL STRATEGY WITH CLUSTERING STRATEGY - LARGEST EIGENVALUE OF LARGEST CONNECTED COMPONENT AND LARGEST CLUSTER.

The clustering with random removal of links has led us to conclude that the largest eigenvalue of the largest cluster can be less or comparable to the largest eigenvalue of the biggest component generated by random links removal. However, other clusters have a significantly smaller largest eigenvalue, which leads to a smaller amount of cleaning necessary to completely remove the worm from the network. Furthermore, if only the largest cluster is infected, only up to 50% of the network will need cleaning.

This paper considers modularity to be the partitioning algorithm, but there exists a large number of partitioning algorithms that try to optimize different variables. The investigation of how different clustering algorithms affect the epidemic dynamics stands on the agenda for future work.

ACKNOWLEDGMENTS

This research was supported by the European Commission, under Grant No FP7-224619 (the ResumeNet project) and by Next Generation Infrastructures (Bsik).

REFERENCES

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd int. workshop on Link discovery*, pages 36–43. ACM Press, 2005.

[2] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[3] L. Briesemeister, P. Lincoln, and P. Porras. Epidemic profiles and defense of scale-free networks. In *WORM '03: Proceedings of the 2003 ACM workshop on Rapid malcode*, pages 67–75, New York, NY, USA, 2003. ACM.

[4] Microsoft Malware Protection Center. Technical details on worm:win32/koobface.gen!d. Technical report, Microsoft, 2009.

[5] D. Chakrabarti, J. Leskovec, C. Faloutsos, S. Madden, C. Guestrin, and M. Faloutsos. Information survival threshold in sensor and p2p networks. In *INFOCOM*, pages 1316–1324. IEEE, 2007.

[6] Y. Chen, G. Paul, S. Havlin, F. Liljeros, and H. E. Stanley. Finding a better immunization strategy. *P.R.L.*, 101(5):058701, 2008.

[7] M. Ciubotariu. Worm js.yamanneram. Technical report, Symantec, 2006.

[8] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[9] D. M. Cvetkovic, M. Doob, and H. Sachs. *Spectra of Graphs: Theory and Applications, 3rd Revised and Enlarged Edition*. Vch Verlagsgesellschaft Mbh, December 1998.

[10] D. J. Daley and J. Gani. *Epidemic Modelling: An Introduction (Cambridge Studies in Mathematical Biology)*. Cambridge University Press, May 2001.

[11] M. Draief, A. Ganesh, and L. Massoulié. Thresholds for virus spread on networks. In *valuetools '06: Proceedings of the 1st international conference on Performance evaluation methodolgies and tools*, page 51, New York, NY, USA, 2006. ACM.

[12] P. Erdos and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

[13] P. T. Eugster, R. Guerraoui, A. Kermarrec, and L. Massouli. From epidemics to distributed computing. *IEEE Computer*, 37:60–67, 2004.

[14] A. J. Ganesh, L. Massouli, and D. F. Towsley. The effect of network topology on the spread of epidemics. In *INFOCOM*, pages 1455–1466. IEEE, 2005.

[15] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, 2002.

[16] D. Michael Goedecke, Georgiy V. Bobashev, and Feng Yu. A stochastic equation-based model of the value of international air-travel restrictions for controlling pandemic flu. In *WSC '07: Proceedings of the 39th conference on Winter simulation*, pages 1538–1542, Piscataway, NJ, USA, 2007. IEEE Press.

[17] E. M. Jin, M. Girvan, and M. E. J. Newman. Structure of growing social networks. *Physical Review E*, 64(4):046132+, 2001.

[18] E.G. Coffman Jr., Z. Ge, V. Misra, and D. Towsley. Network resilience: Exploring cascading failures within bgp. In *Allerton Conference on Communication, Control and Computing*, October 2002.

[19] S. Kamkar. Technical explanation of the myspace worm. Technical report, http://namb.la, 2005.

[20] A. Lelli. Worm js.twettir. Technical report, symantec, 2009.

[21] Jose Marcelino and Marcus Kaiser. Reducing influenza spreading over the airline network. *PLoS Currents Influenza*, Aug 2009.

[22] J. M.Epstein, D. M. Goedecke, F. Yu, R. J.Morris, D. K. Wagener, and G. V.Bobashev. Controlling pandemic flu: The value of international air travel restrictions. *PLoS ONE*, 2(5):e401, 2007.

[23] P. Van Mieghem and J. Omic. In-homogeneous virus spread in networks. Technical report, Technical report 20080801, TUDelft, 2008.

[24] P. Van Mieghem, J. Omic, and R. E. Kooij. Virus spread in networks. *IEEE/ACM Trans. Netw.*, 17(1):1–14, 2009.

[25] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, June 2006.

[26] M. E. J. Newman, S. Forrest, and J. Balthrop. Email networks and the spread of computer viruses. *Phys. Rev. E*, 66(3):035101, Sep 2002.

[27] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, Feb 2004.

[28] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3203, Apr 2001.

[29] Y. Wang, D. Chakrabarti, C. Faloutsos, and C. Wang. Epidemic spreading in real networks: An eigenvalue viewpoint. In *In SRDS*, pages 25–34, 2003.