Anonymity in the wild: Mixes on unstructured networks

Shishir Nagaraja

Computer Laboratory
JJ Thomson Avenue, Cambridge CB3 0FD, UK
`shishir.nagaraja@cl.cam.ac.uk`

Abstract. As decentralized computing scenarios get ever more popular, unstructured topologies are natural candidates to consider running mix networks upon. We consider mix network topologies where mixes are placed on the nodes of an unstructured network, such as social networks and scale-free random networks. We explore the efficiency and traffic analysis resistance properties of mix networks based on unstructured topologies as opposed to theoretically optimal structured topologies, under high latency conditions. We consider a mix of directed and undirected network models, as well as one real world case study – the LiveJournal friendship network topology. Our analysis indicates that mix-networks based on scale-free and small-world topologies have, firstly, mix-route lengths that are roughly comparable to those in expander graphs; second, that compromise of the most central nodes has little effect on anonymization properties, and third, batch sizes required for warding off intersection attacks need to be an order of magnitude higher in unstructured networks in comparison with expander graph topologies.

1 Introduction

As governments pursue large scale surveillance and censorship programs, anonymity in online communication mechanisms is an increasingly important requirement. Anonymous communications are also useful in building resistance against a global passive adversary who can subject the targets to traffic analysis. Often, an attacker will try to destabilize a network by building a dossier of the most central nodes, and attacking ones on the top of the list. Traffic analysis of inter-node communication offers basic tools to collect necessary intelligence in order to plan an attack.

Seminal work by Chaum [Cha81] introduced mix networks as a technique to provide anonymous communications where messages are relayed through a sequence of intermediate nodes called mixes, to make the task of tracing them through the network as difficult as possible. The essential idea is to make the inputs of each mix bit-wise unlinkable to its outputs.

Anonymity research conducted since, can be classified into low-latency or real time systems primarily for Internet browsing such as onion routing [STRL00] and high-latency or non-real time systems such as mixminion [DDM03].

The topology of a mix network plays an important role in its efficiency and traffic analysis resistance properties. The mainstream design paradigm that has emerged so far is that of structured network topologies based on regular graphs. The theory is that such topologies are amenable to theoretical analysis that proves they have optimal expansion properties. This leads to a mix network design that is highly efficient and resistant to traffic analysis. Examples are onion-routing systems such as TOR [DMS04]that use a complete graph topology, where a mix can contact every other mix in the network. While such models are theoretically elegant, the assumption that every node in the network is equally resourced (as regular graphs necessitate) to handle network traffic loads is their main drawback.

An alternate paradigm is topology based on unstructured networks, such as those inspired from social networks. The argument in their favor being that the incentive to carry traffic is clear and simple - friends carry each-others traffic. Moreover, no additional resources go into constructing an overlay network since the pre-existing topology is used by the mix network as well, which works well for power constrained environments such as adhoc networks and sensor networks. Legal considerations play an important role too. It is not enough to merely have a large number of mixes. When hassled by legal requests (such as a subpoena to hand-over mix server logs to the police), a mix-network where friends route each others traffic, is likely to have a higher proportion of servers in operation, as opposed to a synthetic network.

A comparison between the two paradigms needs to address mix-network efficiency, resilience to corrupt nodes and the loss of anonymity from statistical disclosure attacks.

In this paper we analyze various types of unstructured networks, especially social networks and evaluate their suitability as mix topologies. We discuss the reasons behind using social networks to route mix traffic and we analyze the suitability of various types of model networks to routing mix traffic and offer a comparison between them. We also analyze the theoretical bounds on anonymity such networks can provide in terms of mixing speed and resistance to traffic analysis. We apply concepts from spectral graph theory to derive the route length necessary to provide maximal anonymity.

This paper is organized as follows: Section 3 discusses the various topologies used in our analysis. Section 4, lays out the evaluation framework to measure the traffic analysis resistance of various topologies. Section 5 discusses the application of the framework to various topologies and the results obtained. Finally, we offer our conclusions in section 6.

2 Related work

Danezis [Dan03] explored the anonymity provided by expander graph topologies, this is one of the main sources of inspiration for our work. He established the thoretical bounds of anonymity for expander graphs, and also showed that they were optimal.

Borisov [Bor05] analyzes anonymous communications over a De Bruijn graph topology overlay network. He analyzes the deBruijn graph topology and comments on their successful mixing capabilities.

3 Network models

In this section we give a brief introduction to the network models we wish to analyze as candidates for mix network topologies.

3.1 Erdös-Rényi model of random networks

On the earliest models for heterogeneous networks is the Erdös-Rényi (ER)model [ER59]. Although seldom found in real world networks, their use has been popularised by the work of Eschenauer and Gligor [EG02] is designing a key management scheme for sensor networks.

Here, we start from N vertices without any edges. Subsequently, edges connecting two randomly chosen vertices are added as the result of a Bernoulli trial, with a parameter p. It generates random networks with no particular structural bias. The average degree $\langle k \rangle = 2L/N$ where L is the total number of edges, can also be used as a control parameter. ER model networks have a logarithmically increasing l, a normal degree distribution, and a clustering coefficient close to zero.

3.2 Scale-free networks with linear preferential attachment

A number of popular peer-to-peer systems are found to have heterogeneous topologies with heavy tailed degree distributions. The work of Ripeanu [RFI02] shows that two popular systems, Gnutella [KM02] and Freenet [CSWH00], have power-law degree distributions.

A variable X is said to follow a heavy tail distribution if $Pr[X > x] \sim x^{-k} L(x)$ where $k \in \Re^{+}$ and $L(x)$ is a slowly varying function so that $\lim_{x \to \infty} \frac{L(tx)}{L(x)} \to 1$. A power-law distribution is simply a variation of the above where one studies $Pr[X = x] \sim x^{-(k+1)} = x^{-\alpha}$. The degree of a node is the number of links it has to other nodes in the network. If the degree distribution of a network follows a power-law distribution it is known as a scale-free network. The power-law in the degree or link distribution reflects the presence of central individuals who interact with many others on a continual basis and play a key role in relaying information.

We denote a scale-free network generated by preferential attachment, by $G_{m,N}(V, E)$ where m is the number of initial nodes created at time=t_0 and N is the total number of nodes in the network. At every time step $t_i, i \geq 0$, one node is added to the network. For every node v added, we create m edges from the v to existing nodes in the network according to the following linear preferential attachment function due to Barabasi and Albert [AB02]:

$$Pr[(v, i)] = k_i / \sum_j k_j$$

where k_i is the degree of node i. We continue until $|V| = N$.

3.3 Scale-free random graph topology

An alternate way of constructing a large scale-free network is to create a network with a given power-law degree sequence that is random in all other aspects. Aiello et.al. [ACL00] propose such a random graph model inspired by massive AT&T call graphs, with two parameters α and β. Where, α gives the fraction of nodes with degree 1 and β defines the exponent of the power-law function. Then, if y be the number of vertices of degree $x > 0$, x and y satisfy $log(y) = \alpha - \beta log(x)$.

3.4 Klienberg-Watts-Strogatz(KWS) small world topology

Our next network model is inspired by the network of social contacts. It is well known that any two people are linked by a chain of half a dozen others who are pairwise acquainted – known as the 'small-world' phenomenon. This idea was popularised by Milgram in the 60s [Mil67].

The KWS graph topology models a small world network that encapsulates the following: a network rich in local connections, with a few long range connections. The network generation starts from a N by N lattice

each point representing an individual in a social network. The lattice distance $d((i,j),(k,l)) = |k-i| + |l-j|$. For a parameter p, every node u has a directed link to every other node v within $d(u,v) \leq p$. For parameters q and r, we construct q long range directed links from u to a node v with a probability distribution $[Pr(u,v)] = \frac{(d(u,v))^{(-r)}}{\sum_v (d(u,v))^{(-r)}}$.

Low r values means long-range connections, whereas higher values lead to preferential connections in the vicinity of u.

3.5 LiveJournal (LJ)

In order to test our ideas on a real world unstructured network, we turned to a large-scale social network called LiveJournal (LJ). LiveJournal is a social networking and blogging site with several million members and a large collection of user defined communities. LiveJournal allows members to maintain journals, individual and group blogs, and – most importantly for our study here – it allows people to declare which other members are their friends. Using a web crawler called touchgraph (http://www.touchgraph.com), we traced the LJ network to the online friendship network. The snapshot of the network we use in our analysis has 3,746,240 nodes and 27,430,000 edges.

A mix server bundled along with a future LiveJournal client acts as the basis of mix deployment. Mix circuits are built on top of the social network topology.

3.6 Expander graphs

Danezis [Dan03] previously analyzed the use of expander graph topologies to construct mix networks. Expanders are well known to have excellent expansion properties. We include this as a baseline comparison against theoretical structured topologies. An expander graph $G_{N,D}$ has a homogeneous topology with N nodes each with a degree D.

4 Evaluation framework for measuring traffic analysis resistance

Before we set out the evaluation framework, we first clarify what we mean by "anonymity" in this paper. The focus of this work is on message receiver anonymity [SD02]: given a message, the attacker should not be able to determine who sent it to whom, leading to both sender and receiver anonymity requirements. Sender anonymity is determined by the probability that a specific node is the originator of a given message. Receiver

anonymity, also an important requirement in a number real world situations, is the probability that a specific node is the recipient of a given message whose sender is known. There are other definitions such as relationship anonymity defined by Pfitzmann et. al. [PH00]. We also note there that the evaluation framework is the contribution of Danezis [Dan03].

The objective of our analysis is to determine how the topology of a mix network affects the amount of effort on the attacker's part to uniquely identify communication endpoints using traffic analysis attacks alone. The effectiveness of such attacks depends heavily on the topology of the underlying network. If the attacker is not able to reduce anonymity beyond his or her initial knowledge then the mix network is said to be resistant to traffic analysis attacks under the given threat model.

The attacker might also employ side channel analysis on the endpoints before the data enters the mix network, we do not consider such attacks here. Side channel information might be timestamps or other information related to the protocol or mechanism in use. Attacks using such information can be used to link messages to the communication end-points, and are known as *traffic confirmation attacks* [RSG98], their effectiveness depends on the mixes' batching and flushing strategy.

4.1 Threat Model

Throughout this paper we consider the adversarial context of a global passive adversary.

4.2 Measuring anonymity

There are several ways one can express the anonymity a system provides. In our analysis we use a quantitative method due to Serjantov and Danezis [SD02], based on the following definition: "Anonymity of a system may be defined as the amount of information the attacker is missing to uniquely identify an actor's link to an action". In information theoretic terms, the anonymity of the system $\mathcal{A}$, is the entropy $\mathcal{E}$, of the probability distribution over all the actors α_i, that they committed a specific action.

$$\mathcal{A} = \mathcal{E}[\alpha_i] = -\sum_i Pr[\alpha_i] log_2 Pr[\alpha_i] \qquad (1)$$

This gives the number of bits of information, with a negative sign, that the attacker is missing before they can uniquely identify a sender or a receiver.

4.3 Modeling mix route selection

In order to understand the maximal anonymity provided by a mix network we use Markov chains to model the route selection process, as they closely match the way mixes are selected to form a mix route.

The process of selecting a mix route of length k by selecting k random nodes in the mix network, is equivalent to first selecting a random mix node, and, then a random neighbour of the first mix, repeating this process $k - 2$ times. Hence we may model the route selection process as a random walk on the underlying graph, with the various states of the Markov chain process being the mix nodes of the network.

4.4 Measuring mix network efficiency

Receiver anonymity

In analyzing the receiver anonymity provided by a particular network topology we need to examine the probability that a specific message is at a particular node at a certain time. In order to link the sender and the receiver to a particular message, the attacker must retrace the steps taken by the message through the mix network starting from the receiver. Let the mix network be an undirected graph $G(V, E)$. If messages m_{ij} are inserted at node i destined for j, then for a message m_x^t at node x at time t, the attacker must link m_x^t to m_{ij}. Note that m_x^t might either be in the edge or the core of the mix network.

Applying the above mentioned information theoretic metric we have:

$$\mathcal{A} = \mathcal{E}(p_{ij})$$

where $p_{ij} = Pr[m_x^t \; is \; m_{ij}]$ is the probability distribution over all the nodes in V.

Suppose a message is inserted into the mix network through a randomly chosen node. Then after an infinite number of steps, the probability that the message is present on any randomly chosen node in the network is given by stationary distribution of the Markov chain π. Let $q^{(0)}$ be the initial probability distribution describing the node on which message m is introduced into the mix network, this is equivalent to the distribution of input load across the nodes in network. $q^{(t)}$ then, is the probability distribution of the node on which the message is present after t steps. (this is also known as the state probability vector of the Markov chain at time $t \geq 0$). With increasing t one would like to see that $q^{(t)}$ merges with π. The rate at which this takes place is known as the *convergence rate* of

the Markov chain, and the difference itself is called the *relative point-wise distance* defined as:

$$\Delta(t) = max_i \frac{|q_i^t - \pi_i|}{\pi_i} \quad (2)$$

The smaller the relative point-wise distance, faster the convergence, and more efficient the mix network. It is now easy to see that the maximum receiver anonymity $Pr[x = receiver|y = sender]$ the network can provide is the entropy of the stationary distribution of the chain.

$$\mathcal{A}_{network} = \mathcal{E}(\pi) \quad (3)$$

.

When P is the transition matrix of the chain it is well known that P has n real eigen-vectors π_i and n eigenvalues λ_i [Wes01].

By using the relation $q^{(t)} = q^{(0)}P^{(t)}$, we calculate the probability distribution of a message being on a node after having transited a mix route of length t.

Sender anonymity

Next, we consider the probability distribution of potential originators of a given message recipient. This may also be modeled by a Markovian random walk. For a destination node y, consider all random walks terminating at y. In order to achieve maximal sender anonymity, all these walks must be long enough for the respective state probability vector to converge with the stationary distribution. Since this applies equally to all sender nodes in the network [Bor07], the sender anonymity is given by:

$$Pr[X = x|y] = \frac{1}{N = |V|}$$

.

Hence, both maximal sender and receiver anonymity are achieved when the random walk reaches convergence.

Also, the stationary distribution vector gives the normalized fraction of traffic load on each mix [Dan07].

4.5   Compromised mixes

Suppose a subset of mixes are taken over by an adversary. Then a compromised mix route is defined as a mix circuit that is solely composed of compromised mix nodes. Then, what is the probability that a randomly chosen mix route is compromised?

A network topology with poor expansion properties (or lower *eigen-value gap* $\epsilon = 1 - \lambda_2$) tends to have relatively 'localized' mix routes, so that, given the first mix of a route, there exists a subset of mixes within the network that have a higher chance of being on the route than others.

The spectral theory of graphs lends us a few tools, namely chernoff bounds, in quantifying this risk. Suppose S is the set of subverted nodes, and π_S the corresponding probability mass of the stationary distribution π. The upper bound of the probability that a mix route (random walk) of length t goes through t_S nodes of S is given by Gilbert [Gil98]: $Pr[t_A = t] \leq \left(1 + \frac{(1-\pi(A))\epsilon}{10}\right) e^{-t\frac{(1-\pi(A))^2\epsilon}{20}}$. However as Danezis [Dan03] notes, given that this probability exponentially decreases with increase in t, a small increase in route length will successfully mitigate this risk.

What is more relevant in the context of unstructured networks, is the presence of 'hub' nodes and 'weak-ties'. Hubs [New03b] are special nodes that owing to their position in the network topology handle large amounts of traffic. Similarly, weak-ties [Gra73] are edges responsible for significantly reducing average path-lengths in networks of tightly knit communities such as social networks. The risk of compromised mix routes is significantly higher in a topology where hubs only connect to other hubs, and handle most of the network traffic. If an attacker can locate and strategically target mix nodes that also play the role of a hub, then the percentage of mix routes under risk can be significant. This property is known as assortativity [New03a], defined as the affinity of a node to link to others that are similar or different in some way.

Hence, we simulated a large number of random walks for various topologies presented in section 3, of different lengths, and make a recommendation on the route length to mitigate this risk in section 5.1.

4.6   Intersection attacks

The term *intersection attack* was introduced by Berthold et.al. [BPS00]. These attacks involve the detection of the preferential use of a mix route. If for some reason, a sender under attack sends more traffic along a specific route much more often than other routes, then a simple intersection attack is carried out by intersecting the set of possible next-hop mixes of every mix with the set of possible next-hop destinations of previous messages. The the actual path of a message will then become apparent unless the network has countermeasures against observability.

If each link from a mix node is used to flush messages to its neighbours, then the potential for the simplest of intersection attacks can be

greatly reduced [KAP02]. So, for a given node i, we wish to calculate the probability that any out going link remains unused during a flushing cycle. If each mix node receives b messages per batch, then each of these will appear on a particular outgoing link j with a binomial probability distribution $p_i = 1/deg_i$. Danezis [Dan03] then calculates the volume of incoming traffic required so that the probability of any out going link being unused is negligible.

$$b = \frac{9}{f^2} \left(\frac{1 - p_i}{p_i} \right) \tag{4}$$

where f is the percentage deviation of traffic output on a particular link of i in a given flushing cycle from the mean traffic output.

Combining this with p_{min}, the probability associated with the highest degree node in the mix network, we can derive the amount of genuine traffic to be mixed together.

The prevention of basic intersection attacks as a system design criteria can be traced back to the work of Reiter and Rubin [RR98].

5 Results and Discussion

5.1 Simulation parameters

In all the synthetically generated networks we considered, we $N \approx 5000$ nodes. The parameters used for each of them are listed below.

We model scale-free networks with linear preferential attachment with m links per node and average node degree $\langle d \rangle$; $2 \leq m \leq 7$ and $4 \leq \langle d \rangle \leq 14$.

Next we model scale-free random networks which have a scale-free degree sequence but which are random in all other respects. Generated with parameters $\alpha = 0.25$, $\beta = 0.25$ and Average node degree $4 \leq \langle d \rangle \leq 14$. See section 3.3 for an explanation of α and β

Klienberg-Watts-Strogatz model of directed social network ties is analyzed next, generated with parameters r, the lattice radius within which each node creates direct links to all its neighbors. q is the number of weak ties. We used $1 \leq r \leq 4$ and $2 \leq q \leq 10$.

Our next network is based on our primary source data, obtained by web-crawling the LiveJournal site. The snapshot of the network we use in our analysis has 3,746,240 nodes and 27,430,000 edges.

Finally we analyze two theoretical topologies, one degree heterogeneous and the other degree homogeneous, to offer a baseline comparison against ER graph and constant expander graph topologies.

The ER graph is created with each edge formation as the result of a Poisson distribution of $p = 0.0028$ with $\langle d \rangle = 14$.

The constant expander graph is created with each node having $D = 14$ edges. Motwani et.al. [MR95] prove a relation between the second eigenvalue λ_2 of the transition matrix of a constant expander graph and the degree D of a node $\lambda_2 \geq \frac{2\sqrt{D-1}}{D}$. We can then use the result of Sinclair [Sin93] connecting λ_2, random walk length t and convergence rate $\Delta(t)$, namely $\Delta(t) \leq \frac{\lambda_2^t}{\min_{i \in V} \pi_i}$. For $D = 14$, we have a constant expander graph with theoretical minimum second eigen-value of $\lambda_2 \geq 0.5527708$, converging to maximal anonymity state in approximately 4 steps. This forms the baseline against which we compare all the other topologies.

5.2 Efficiency

We can now comment on the efficiency and recommended mix route lengths for various network topologies by comparing them to our baselines.

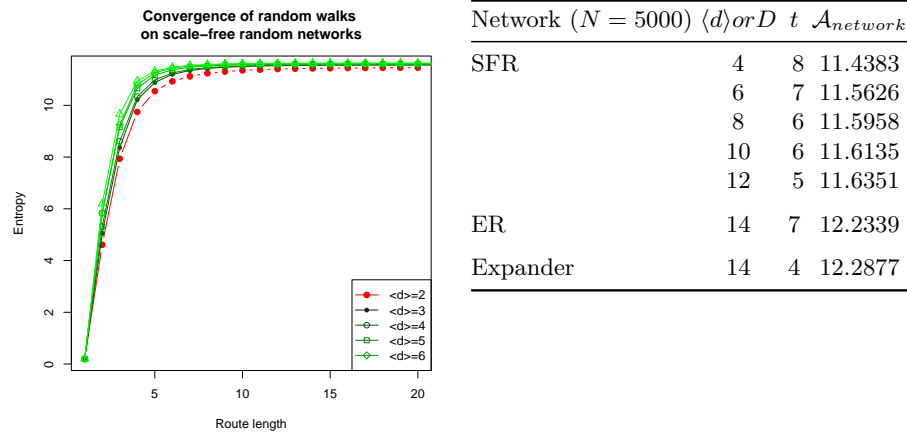| Network ($N = 5000$)	$\langle d \rangle$ or D	t	$\mathcal{A}_{network}$
SFR	4	8	11.4383
	6	7	11.5626
	8	6	11.5958
	10	6	11.6135
	12	5	11.6351
ER	14	7	12.2339
Expander	14	4	12.2877

Fig. 1: Convergence rates: Efficiency and maximal receiver anonymity for Scale-free random, ER and Constant expander graph topologies

The efficiency of mix topologies based on a scale-free random networks is shown in Figure 1. It plots the anonymity achieved against increasing random walk lengths. Maximal receiver anonymity is calculated using equation 3 is the entropy of the probability distribution of the chain at convergence, while maximal sender anonymity is $\frac{1}{N}$.
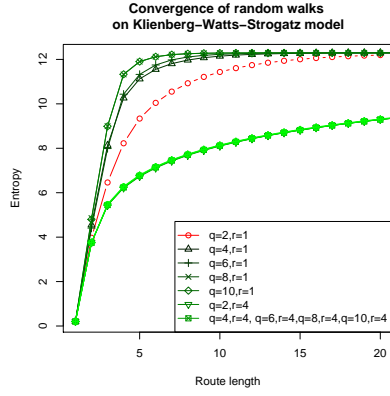
Our calculations show that maximal anonymity is reached in just 6 steps in the medium density case $\langle d \rangle \geq 4$, as opposed to 4 steps in to 4 steps in an expander graph topology. It turns out that social collaboration networks [New01a,New01c,New01b] with scale-free characteristics have average degrees in the range of $4 \leq \langle d \rangle \leq 18$. This suggests, firstly, that efficient mix networks can be designed using scale-free random networks, and second, that mildly denser scale-free networks are more suitable for building mix networks than sparser ones.

While this is an encouraging initial result, it is important to strike a note of caution. Scale-free random graphs only model the scale-free aspect of degree distribution, while being random in every other way. However most real world unstructured networks have several other non-random characteristics apart from their degree distributions.
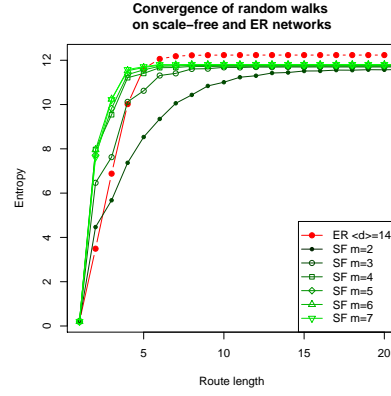
A number of real world unstructured networks are not scale-free, hence we included the Klienberg-Watts-Strogatz(KWS) network topology, as it explicitly models the presence of weak ties in a network. We experiment with a number of parameter configurations; selecting $r = 1$ and $r = 4$ to model low and high richness in local links or 'strong ties' between nodes; and $2 \leq q \leq 10$ the number of short cuts or 'weak ties', between mix nodes. Figure 2-a plots mix-route length vs mix network anonymity, for the KWS topology. When the topology is poor in local links, it seems to converge in 7 to 8 hops, given enough short cuts. However, if the network invests a large amount of resources into local connections forming relatively tightly knit communities, then regardless of the amount of shortcuts, convergence is not achieved until 62 hops!

Our final model network topology is the scale-free network based on linear preferential attachment, which has attracted much attention in the complex networks literature. This topology models a scale-free network where hubs are connected to other hubs, a pattern that is repeatedly observed in many real world scale-free networks. The parameter m controls graph sparsity, random walk and convergence results are shown in figure 2-b. Our simulations show that while very sparse topologies converge in 10 to 15 hops, topologies that are relatively dense converge within 6 or so hops, this is comparable to the optimal 4 hops of a constant expander graph.

Next, we considered our primary data source the LiveJournal graph with a little less than 4 million nodes. Figure 3 shows the convergence rate of mix routes, which we note converges to the stationary distribution in around 11 hops. While this seems a high number in comparison to expander graphs (converging in 4 hops), we also note that the entropy

(a) Klienberg-Watts-Strogatz model

(b) Scale-free network with preferential attachment
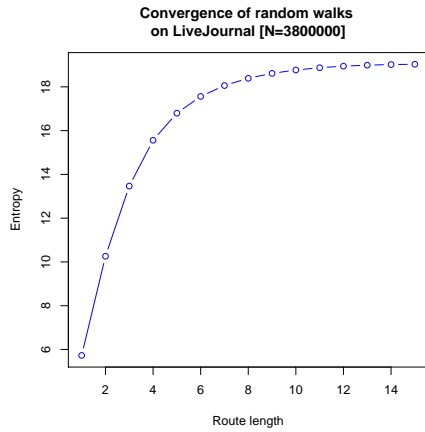
Fig. 2: Mean entropy vs mix-route length

Fig. 3: Mean entropy vs mix-route length for LiveJournal

Network	m	t	$\mathcal{A}_{network}$
SF	2	15	11.5852
	3	10	11.6961
	4	6	11.7293
	5	6	11.7687
	6	6	11.7953
	7	6	11.8090
KWS	$q = 2, p = 1$	11	12.2945
	$q = 10, p = 1$	5	12.2939
	$q = 2, p = 4$	63	11.6440
	$q = 10, p = 4$	63	11.6380
ER	$\langle d \rangle = 14$	7	12.2339
Expander	$D = 14$	4	12.2877

Fig. 4: Convergence rates: Efficiency and maximal receiver anonymity for linear Scale-free, KWS, ER and constant expander graph topologies

achieved by the random walk in 4 hops in LJ is $\mathcal{A}_{LJ}^4 = 15.56$. To obtain the equivalent on an expander graph topology we would only need $2^{15.56}$ or 48309 nodes. On the face of it the design decision seems really simple, to go for a structured expander graph topology. We argue a different view: A successful mix network design must also consider liability management issues arising from running a mix. Considering that aspect, topology links backed up by social capital are likely to be more robust than those of an optimal topology, but where nodes quickly buckle under legal pressure. We propose, running mixes on the nodes of the LJ topology bundled along with a future LiveJournal client. Nodes only allow incoming traffic from their neighbors, and will only direct outgoing traffic to their neighbors.

In this context, an interesting question is why nodes would process traffic that didn't originate from their neighbors, and especially so in the face of legal hassles?

We offer the following reasoning: Humans making decisions on whether or not to run a mix server, will have to consider the following costs. They benefit in the long term, from processing traffic for unknown nodes in order to generate a diverse user base, the need for which is well illustrated in Dingledine and Mathewson [DM06]. However this only holds if other mixes cooperate accordingly. Then there is the immediate social benefit of having processed traffic for your friends. The success of the system then depends on the extent to which individual nodes perceive the costs of litigation pressure to be less than the total of immediate social benefit and the long term benefit of a diverse user base. Psychology studies tell us that humans involved in taking security decisions weigh short and long term benefits differently. It should also be interesting to investigate whether the idea of running a mix to primarily process traffic for your friends is an effective tool for seeding indirect reciprocity in a mix network where cooperation flourishes.

5.3   Compromised mix nodes

As explained in our evaluation framework, compromised nodes can lead to compromised routes. This presents a special challenge in unstructured networks where π_A, the probability mass of the stationary distribution π, corresponding to set of compromised nodes A, can be significant for topological reasons.

To measure the robustness to nodes being strategically compromised by an attacker, we simulated 100000 random walks of different lengths, for each of our network topologies, in the range indicated by efficiency considerations of the previous section $3 \leq t \leq 6$, and measured the fraction

that passed through compromised nodes. The set of compromised nodes is chosen to consist of the nodes with the highest degrees in the network. In each case, for mix routes greater than 4 hops the probability of existence of a compromised mix route is negligible. Fig 5 in the appendix confirms that the threat of mix route compromise can be successfully reduced by suitably increasing the mix-route length.

5.4   Intersection attacks

Using equation 4 we consider the required batch sizes for a threshold mix, so that the traffic output on any link in the mix network does not deviate by more than 5% from the mean traffic output on that link. For $f = 5$ we calculate the number of messages that must be received in each mixing cycle in table 1.

Network	$\langle d \rangle$	p_{min}	Batch size
SFR	4	0.0344	10.08
	6	0.0222	15.84
	8	0.0243	14.4
	10	0.0192	18.36
	12	0.0135	26.28
	14	0.0125	28.44
KWS	27 $(q = 1, r = 1)$	0.0294	11.88
	43 $(q = 10, r = 1)$	0.0169	20.88
	26 $(q = 1, r = 4)$	0.0333	10.44
	28 $(q = 10, r = 4)$	0.0294	11.88
SF-linear	4	0.0048	74.16
	6	0.0048	74.16
	8	0.0041	86.04
	10	0.0038	93.6
	12	0.0037	96.12
	14	0.0031	112.32
LJ	7.3221	0.00857	41.64
ER	14	0.0333	10.44
Expander	14	0.0714	4.68

Table 1: Batch sizes required to prevent intersection attacks

From table 1 it is clear that scale-free random networks and KWS both require a batch size that is 4-5 times that of expander graphs. Whether social networks can produce enough 'chatter' to feed genuine traffic into the mix network is an open question.

Our theoretical base line of ER network topology does slightly better at a little over twice that. More significantly, the LJ network has a batch size of almost 9 times the required batch size for expander graphs. Scale-free networks with linear preferential attachment are the worst performing, requiring a batch size almost 20 times larger than expanders. We think that the exceptionally high value of batch size in LJ network is due to its size of four million or so nodes. While does not mean that LJ is inherently unsuitable as a mix network topology, but it certainly indicates a scalability limit with the deployment of mixes on LJ nodes, as proposed earlier.

6   Conclusions

We have analyzed a comprehensive set of network topologies from the perspective of efficiency, maximal anonymity, compromised nodes and simple intersection attacks in comparison with (provably optimal) expander graphs.

To the standard threat model of the global passive adversary, we have added real world issues such as liability management and the need for clear incentives for carrying traffic under the pressure of legal threats, and discussed our simulation results in this context.

We have considered topologies with two important characteristics found in the empirical studies of large-scale unstructured networks: scale-freeness (scale-free random graph) and the small-world property (Klienberg-Watts-Strogatz (KWS) graph). In both the topologies, we can recommend mix route lengths for achieving 95% of maximal anonymity, that is only a few hops larger than the optimal route length found in expander graph topologies. Currently deployed mix networks such as TOR have around 540 volunteers. To increase the scale of such mix deployments the Internet, we believe the way forward (for high latency systems only) is to use online social networks. The minimum mix route must have three mixes to allow sender and receiver anonymity. For this length, a mix network constructed by placing mixes on the nodes of a social network such as LiveJournal can achieve far higher maximal anonymity as per the entropy metric we have used. We argue that including network incentives within a framework does not allow the construction of structured overlay mix topologies that can robustly withstand the threat of legal action. By moving to social networks, we make a start on tapping the social capital underlying node-node interaction to encourage users to deploy and run mixes with policies that reflect this aspect.

We also found that subverted nodes, either compromised randomly, or by strategic choice, on the basis of their degrees has little effect on the efficiency of a mix network. This is because the route length required to mitigate that risk is less than the recommended length for achieving efficient convergence rates.

We also analyzed scale-free and the small-world topologies for their robustness to attacks based on traffic load patterns observable on their out-going and in-coming links. Both the scale-free random graph topology and the KWS topology turn out to require almost 5 times as much traffic as corresponding expander graph topology. This suggests the need for further tests to see if enough genuine traffic is generated in online social network interaction, to satisfy the minimum batch sizes required for preventing the most basic versions of these attacks.

We conclude that, unstructured networks based on large-scale topologies are indeed very promising, we have outlined the merits and challenges these topologies present to the design of mix networks for anonymous communication.

7 Acknowledgements

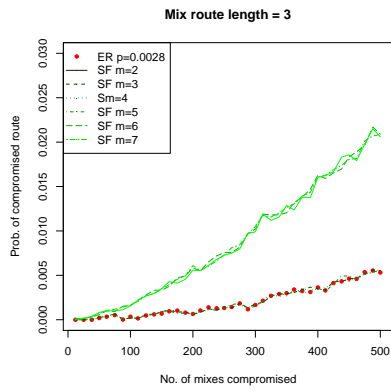## A Mix-route compromise on linear preferential attachment scale-free networks

In this section we sketch a few analytical results concerning mix-route compromise in BA scale-free networks.

Let B be the set of compromised high vertex-order centrality nodes. For a route to be fully compromised, all intermediate nodes must be in B. We then wish to calculate,
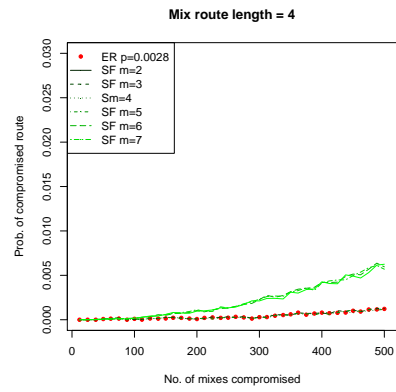
$P(C_l) = [Pr(Random - Walk(v_1....v_l))] \forall v_1...v_l \subseteq B$.

It is straightforward to see that if $l > |B|$ then P(C)=0. In BA scale-free networks, all hubs(high vertex-order) nodes are connected to each other. Hence,
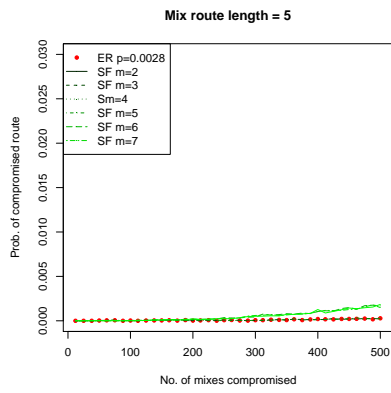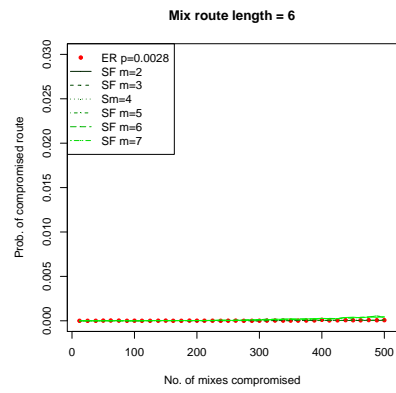
$$P(C) = \frac{|B| - 1}{\prod_{j \in B} k_j}$$

(a) length=3

(b) length=4

(c) length=5

(d) length=6

Fig. 5: Probability of mix-route compromise vs no. of corrupt nodes

B Convergence rate and network size in scale-free random networks

Simulations conducted in this paper have not accounted for the effect of varying network size on the convergence rate of the respective topologies. We address this, by offering a simple conductance based proof that the second eigen-value of a scale-free network is a independent of the network size. See [Ran06] for a review of the conductance based technique as well as others.

We denote a scale-free network generated by preferential attachment, by $G_{m,n}(V, E)$ where m is the number of initial nodes created at time=t_0 and n is the total number of nodes in the network. At every time step $t_i, i \geq 0$, m nodes are added to the networks. For every node added, we create m edges from the node to existing nodes in the network. We continue until $|V| = n$.

Next, there is an intimate relationship between the rate of convergence and a certain structural property called the *conductance* of the underlying graph. Consider a randomly chosen sub-graph S of $G(V, E)$. Suppose a random walk on the graph visits node i $i \in S$. What is the probability that the walk exits S in a single hop. If conductance is small, then a walk would tend to "get stuck" in S, whereas if conductance is large it easily "flows" out of S.

Formally, for $S \subset G$, the *volume* of S is $vol_G(S) = \sum_{u \in S} d_G(u)$, where $d_G(u)$ is the degree of node u. The *cutset* of S, $C_G(S, \overline{S})$, is the multiset of edges with one endpoint in S and the other endpoint in $\overline{S}$. The textbook definition of conductance Φ_G of the graph G is the following:

$$\Phi_G = \min_{S \subset V, vol_G(S) \leq vol_G(V)/2} \frac{|C_G(S, \overline{S})|}{vol_G(S)} \tag{5}$$

[MPS03] prove that the conductance of a scale-free network is a *constant*. Specifically, $\forall m \geq 2$ *and* $c < 2(d-1) - 1$, $\exists \alpha = \alpha(d, c)$ such that

$$\Phi = \frac{\alpha}{m + \alpha} \tag{6}$$

From [Sin93] we have the following bound for λ_2:

$$1 - 2\Phi \leq \lambda_2 \leq 1 - \Phi^2/2 \tag{7}$$

Substituting for Φ from equation 6 in equation 7, it is easy to see that λ_2 is a constant.

References

[AB02]   R. Albert and A. Barabási. Statistical mechanics of complex networks, 2002.

[ACL00]   William Aiello, Fan Chung, and Linyuan Lu. A random graph model for massive graphs. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 171–180, New York, NY, USA, 2000. ACM Press.

[Bor05]   Nikita Borisov. Phd thesis: Anonymous routing in structured peer-to-peer overlays, April 2005.

[Bor07]   Nikita Borisov. Private communication, June 2007.

[BPS00]   Oliver Berthold, Andreas Pfitzmann, and Ronny Standtke. The disadvantages of free MIX routes and how to overcome them. In H. Federrath, editor, *Proceedings of Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability*, pages 30–45. Springer-Verlag, LNCS 2009, July 2000.

[Cha81]   David Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 4(2), February 1981.

[CSWH00]   Ian Clarke, Oskar Sandberg, Brandon Wiley, and Theodore Hong. Freenet: A distributed anonymous information storage and retrieval system. In *Proceedings of Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability*, pages 46–66, July 2000.

[Dan03]   George Danezis. Mix-networks with restricted routes. In Roger Dingledine, editor, *Proceedings of Privacy Enhancing Technologies workshop (PET 2003)*. Springer-Verlag, LNCS 2760, March 2003.

[Dan07]   George Danezis. Private communication, July 2007.

[DDM03]   George Danezis, Roger Dingledine, and Nick Mathewson. Mixminion: Design of a type iii anonymous remailer protocol. In *IEEE Symposium on Security and Privacy*, pages 2–15, 2003.

[DM06]   Roger Dingledine and Nick Mathewson. Anonymity loves company: Usability and the network effect. In *Proceedings of the Fifth Workshop on the Economics of Information Security (WEIS 2006)*, Cambridge, UK, June 2006.

[DMS04]   Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor: The second-generation onion router. In *Proceedings of the 13th USENIX Security Symposium*, August 2004.

[EG02]   Laurent Eschenauer and Virgil D. Gligor. A key-management scheme for distributed sensor networks. In *CCS '02: Proceedings of the 9th ACM conference on Computer and communications security*, pages 41–47, New York, NY, USA, 2002. ACM Press.

[ER59]   P. Erdos and A. Rnyi. On random graphs. *Publicationes Mathemticae (Debrecen)*, 6:290–297, 1959.

[Gil98]   David Gillman. A chernoff bound for random walks on expander graphs. *SIAM J. Comput.*, 27(4):1203–1220, 1998.

[Gra73]   Mark S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[KAP02]   Dogan Kesdogan, Dakshi Agrawal, and Stefan Penz. Limits of anonymity in open environments. In Fabien Petitcolas, editor, *Proceedings of Information Hiding Workshop (IH 2002)*. Springer-Verlag, LNCS 2578, October 2002.

[KM02]   T. Klingberg and R. Manfredi. "gnutella 0.6", June 2002.

[Mil67]   Stanley Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[MPS03]   Milena Mihail, Christos Papadimitriou, and Amin Saberi. On certain connectivity properties of the internet topology. In *FOCS '03: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, page 28, Washington, DC, USA, 2003. IEEE Computer Society.

[MR95]   Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. 1. Cambridge Univ. Press, 1995. Motwani.

[New01a]   M. E. Newman. The structure of scientific collaboration networks. *Proc Natl Acad Sci U S A*, 98(2):404–409, January 2001.

[New01b]   M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(1):016132, Jun 2001.

[New01c]   M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64(1):016131, Jun 2001.

[New03a]   M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67:026126, 2003.

[New03b]   M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[PH00]   Andreas Pfitzmann and Marit Hansen. Anonymity, unobservability, and pseudonymity: A consolidated proposal for terminology. Draft, July 2000.

[Ran06]   Dana Randall. Rapidly mixing markov chains with applications in computer science and physics. *Computing in Science and Engineering*, 8(2):30–41, 2006.

[RFI02]   Matei Ripeanu, Ian Foster, and Adriana Iamnitchi. "mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design". *IEEE Internet Computing Journal*, 6(1), August 2002.

[RR98]   Michael K. Reiter and Aviel D. Rubin. Crowds: anonymity for web transactions. *ACM Trans. Inf. Syst. Secur.*, 1(1):66–92, 1998.

[RSG98]   Michael G. Reed, Paul F. Syverson, and David M. Goldschlag. Anonymous connections and onion routing. *IEEE Journal on Selected Areas in Communications*, 16(4), 1998.

[SD02]   Andrei Serjantov and George Danezis. Towards an information theoretic metric for anonymity. In Paul Syverson and Roger Dingledine, editors, *Privacy Enhancing Technologies*, volume 2482 of *LNCS*, San Francisco, CA, April 2002.

[Sin93]   Alistair Sinclair. *Algorithms for random generation and counting: a Markov chain approach*. Birkhauser Verlag, Basel, Switzerland, Switzerland, 1993.

[STRL00]   Paul Syverson, Gene Tsudik, Michael Reed, and Carl Landwehr. Towards an Analysis of Onion Routing Security. In H. Federrath, editor, *Proceedings of Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability*, pages 96–114. Springer-Verlag, LNCS 2009, July 2000.

[Wes01]   Douglas B West. *Introduction to Graph Theory*. Prentice Hall, second edition, 2001.