The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1066-2243.htm

Received 16 June 2008
Accepted 7 October 2008

Collaborative attack on Internet users' anonymity

Rami Puzis
Deutsche Telekom Laboratories, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Dana Yagil
Department of Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Yuval Elovici
Deutsche Telekom Laboratories, Ben-Gurion University of the Negev, Beer-Sheva, Israel, and

Dan Braha
Department of Management, University of Massachusetts Dartmouth, Dartmouth, Massachusetts, USA

Abstract

Purpose – The purpose of this paper is to model and study the effectiveness of an attack on the anonymity of Internet users by a group of collaborating eavesdroppers.

Design/methodology/approach – The paper is based on an analysis of the Internet topology. The study is based on two methods for choosing nodes that contribute the most to the detection of as many communicating Internet users as possible.

Findings – The paper illustrates that it is possible to compromise the anonymity of many Internet users when eavesdropping on a relatively small number of nodes, even when the most central ones are protected from eavesdropping.

Research limitations/implications – It is assumed that the Internet users under attack are not using any anonymity enhancing technologies, but nodes can be protected from eavesdropping. It proposes a measure of the success of an attack on Internet users' anonymity, for a given deployment of collaborating eavesdroppers in the Internet.

Practical implications – The paper shows that several, and not necessarily the most prominent, collaborating nodes can compromise the anonymity of a considerable portion of Internet users. This study also emphasizes that when trying to completely compromise the anonymity of Internet users, an eavesdroppers' deployment strategy that considers eavesdroppers' collaboration can result in substantial resource saving compared to choosing a set of the most prominent nodes.

Originality/value – The paper proposes a new measure of anonymity level in the network, based on the linkability of the Internet users. This paper is the first to present results of a non-trivial Group Betweenness optimization strategy in large complex networks.

Keywords Internet, User studies, Data security

Paper type Research paper

Internet Research
Vol. 19 No. 1, 2009
pp. 60-77

1066-2243
DOI 10.1108/10662240910927821

1. Introduction

Throughout the world, millions of people are using the Internet at any given moment. Each of these users has his/her own agenda, whether seeking information,

entertainment or communication with another user. Most users leave identifiable tracks when visiting web sites, checking for e-mail, and so on. Such tracks include information provided by the user him/herself, as well as information provided by the communication medium being used (Boyan, 1997). This information may be used to unveil the user's identity and thus to break the anonymity of the communicating parties (Elovici and Glezer, 2003). For example, whenever a user submits simple web requests, eavesdroppers may gain access to a wide range of data, including, amongst others, the user's Internet protocol (IP), the web sites accessed, personal user interests, habits and preferences (Shapira *et al.*, 2005). The Internet is packed with global eavesdroppers, such as governments seeking to identify terrorists, or police forces seeking to identify criminals (Reiter and Rubin, 1998, 1999).

The vulnerability of anonymity derives from the fundamental nature of the IP that is used for routing packets on the Internet (Elovici and Glezer, 2003; Gritzalis, 2004). Pursuant to the IP, each packet has a header, used for routing the packet in the network, and a data payload that carries the data. The header of each IP packet includes the IP address of the machine that sent the packet as well as the IP address of the intended recipient of the packet. Since the header has to be visible to the network and to the observers of the network under normal communication, any eavesdropper situated on the network along the path that a packet travels can easily determine what entities are communicating (source IP, destination IP). Also, any recipient of a packet is able to determine the source directly from received packets (source IP). While IP addresses do not necessarily uniquely identify an individual, it may be possible to link even dynamically assigned IP addresses to an individual, if they access services that require identification with the assigned address, or if records about assignment of addresses to users during a particular period of time are available (for instance, after the September 11 attack, new homeland security related laws in the USA obliged ISPs to keep records of all assigned IP addresses).

In light of the described information flow process, an eavesdropper seeking to launch an attack against anonymity on the Internet will most likely be interested in eavesdropping on every single router on the Internet. This conclusion is based on the assumption that the more information flow covered by an eavesdropper, the greater his or her success in uncovering the identity of communicating parties. By achieving this massive coverage, no information will escape the eavesdropper and he/she may be able to uncover the identity of any communicating entity on the Internet. Nonetheless, this information coverage is both impractical and implausible, since the Internet includes hundreds of thousands of routers. Facing this technical difficulty, an eavesdropper can equip him/herself with a more efficient solution, allowing the goals to be partially achieved by optimally exploiting the existing eavesdropping tools and the computational capability at his/her disposal.

In this study we refer to Internet topology at the autonomous system level (CAIDA, 2008). Autonomous system is a portion of the physical network (routers and links) that is controlled by a single administrative entity such as a network service provider. Routing policies, quality of service, network measurements, etc. within a single autonomous system are in many cases controlled by closely related departments. Different autonomous systems have traffic transfer agreements that allow the users in one autonomous system to communicate with users in another, distant, autonomous

system. An intruder or dishonest employee directly involved with traffic management inside an autonomous system is able to launch an attack on anonymity by linking between communicating parties whose traffic is routed through this autonomous system.

Faloutsos *et al.* (1999), Strogatz (2001), and Vazquez *et al.* (2002) have shown the Internet to be a scale-free network. Barabási and Albert (1999) characterized such networks by a power-law distribution of connectivity degree, meaning that the probability of a node to have k connections to other nodes is proportional to k^{α}. Plots of functions of this form have a long right tail which continues to be non-zero even for very high values of k. Practically, this implies that there are a few very central nodes (with a lot of connections), while most of the network nodes are peripheral. Scale-free networks allow substantial information coverage by eavesdropping on a fairly small number of central nodes (Barthélemy, 2004). Therefore, to preserve the users' privacy, these nodes should be protected from eavesdropping.

Since in our case nodes are autonomous systems, there are no easy ways to protect them. Rules, internal regulations, and tight security can prevent insiders' attacks that may leak sensitive information out of the organization. Nevertheless, if the organization itself decides to eavesdrop or several organizations that own autonomous systems decide to collaborate, they all have chances to successfully launch the attack on Internet users' anonymity.

In this study we show that attackers can attain substantial information coverage even when the most prominent nodes are protected. We evaluate and compare two deployment strategies of a group of collaborating eavesdroppers. In the first strategy the attackers may choose a set of the most prominent non-protected nodes and in the second strategy the nodes are chosen taking into account the mutual effects the nodes have on each other (Puzis *et al*, 2007a, b). Prominence of nodes and groups of nodes is measured by their betweenness centrality (BC) (Freeman, 1977; Everett and Borgatti, 1999). Betweenness is roughly defined as the total fraction of shortest paths between all pairs of nodes that pass through the investigated node (or group of nodes). More extensive explanation of Betweenness is given in the Section 3.

The outline of this paper is as follows: Section 2 provides a review on anonymity in the web and on the BC measure, and its usage in complex networks. In Section 3 we quantify the success of the attack on anonymity when eavesdropping on nodes on the Internet using the BC measure. In Section 4 we quantify the success of the attack on anonymity of a group of collaborating nodes. Section 5 illustrates through simulations the possible achievements of attackers and defenders of users' anonymity. Section 6 presents the conclusions of the study.

2. Related work

Work related to this study includes two different fields, the first is anonymity in the web and the second field is BC measure. Presented below is a brief overview of both fields, which serves as background material for this study.

2.1. Anonymity in the web

The web is based on a client-server model. Users are clients of web sites, which are hosted on servers. Instead of using the terms user, initiator or client, all of the above

will be referred to as sender, and the terms web-server, surfed-site or end-server, will be referred to as recipient. Both sender and recipient communicate using standard protocols, such as TCP/IP, which is widely used in the Internet network.

What then, is anonymity? Pfitzmann and Hansen (2008) defined anonymity as "the state of being not identifiable within a set of subjects, the anonymity set". A sender is identifiable when one receives information that can be linked to the sender, e.g. when the IP address of the machine used by the sender can be linked to him/her. Pfitzmann and Waidner (1986, 1989) described sender and receiver anonymity as the ability to conceal sender and receiver identity from an attacker. They defined unlinkability as the ability to prevent an attacker from linking the actual message sent by the sender to the actual message received by the recipient. Gabber *et al.* (1997) defined the anonymity of a user as a state in which the identity of the user is kept secret; that is, a web site or a coalition of web sites cannot (except with negligible probability) determine the true identity of the user from his/her alias(es).

Attacks on sender or recipient anonymity could be committed by any entity in the path between the sender and the recipient that is capable of listening to some or all the communication flow, which includes bodies such as the local area administrator, the user's Internet service provider, the recipient him/herself, who might wish to reveal the sender's identity and interests, or the sender, who might "learn" about information stored on the recipient's computer. An attack can also be committed by collaborations, consisting of some senders, receivers, and other parties, or variations of these.

Reiter and Rubin (1998, 1999) considered the anonymity properties available to an individual user against four distinct types of attackers:

(1) *Local eavesdropper*. An attacker who can observe all communication to and from the user's computer. The local eavesdropper could, for example, be an eavesdropper on the local area network of the user, such as an administrator monitoring web usage at a local firewall, or the Internet service provider.

(2) *Global eavesdropper*. An attacker who can observe all the communication to and from a sender and a recipient and thus may learn who is communicating with whom and what information is exchanged.

(3) *Recipient (the end server)*. An attacker who receives the sender's requests and sends replies, and by doing so may learn the sender's interests.

(4) *Collaborating entities*. A group of entities (users, servers, etc.) working together in order to analyze and derive information from a communication that is observable by them.

An attacker may use various kinds of attacks in order to expose the sender and the recipient such as those described by Berthold *et al.* (2000), Raymond (2001), or Song and Korba (2002). Some of the attacks are categorized as "general traffic analysis attacks". Traffic analysis attacks pose a problem in regard to the preservation of confidentiality of conversers' identities and of the conversation's timing (Shields and Levine, 2000).

In an attempt to provide some technical solutions to fill the apparent privacy void for computer network information exchange, several network-based privacy-enhancing technologies that provide varying levels of private

communication between parties have been developed in recent years (Elovici and Glezer, 2003; Reiter and Rubin, 1998; Gritzalis, 2004; Claessens *et al.*, 1999; Chaum, 1981; Goldschlag *et al.*, 1999; Freedman and Morris, 2002; Shields and Levine, 2002). The basic concept underlying these solutions for preserving anonymity on the Internet is the hiding of identifiable tracks "leaking" from the communication infrastructure, by creating an anonymous channel between a sender and a recipient. The anonymous channel is based on dedicated servers or user agents, or a combination of both. The goal of these approaches is to protect users against traffic analysis attacks, where an adversary can match a message sender with the receiver. Some of these methods can be implemented by network service providers supplying a revocable anonymity service to Internet users (Claessens *et al.*, 2003). However, the methods mentioned above have not proven to be completely immune to traffic analysis attacks or require a trusted entity that will provide anonymity services.

Anonymity level is usually a property of a system for anonymous communication and indicates the ability of an attacker to identify the senders and/or recipients of the communications. In order to evaluate the performance of a growing number of different anonymity services, several metrics have been proposed for quantifying the level of anonymity. Most of these metrics are based on the size of the anonymity set (within which the sender and/or the recipient cannot be identified), on distribution of the probability that a particular user is the sender or recipient of the communication, or both. Serjantov and Danezis (2002) and Díaz *et al.* (2003) independently proposed a measure of anonymity based on Shanon's entropy. Edman *et al.* (2007) proposed a system-wide metric based on the permanence of a matrix. In this paper we define a global metric for unlinkability of Internet users given a set of collaborating eavesdroppers.

In this study, we assume that the Internet users are not using any privacy-enhancing technologies and that their communication is exposed. However, authorities may assist users to maintain their privacy by tightening the security of the most prominent nodes in the network and preventing leakage of sensitive information from the responsible organization. Collaborating eavesdroppers resemble distributed collaborative intrusion detection systems that sample traffic in order to detect the outbreak of Internet worms in their early stages (Yegneswaran *et al.*, 2004; Cai *et al.*, 2005; Shmatikov and Wang, 2007). Despite the apparent differences in the target and architecture of collaborative intrusion detection and collaborating eavesdroppers discussed in this paper, there are many similarities in methods that can be utilized by both. Elements of both systems must share information gathered during their operation in order to increase effectiveness of the whole system. Both systems utilize mainly passive monitoring. The primary characteristic of this technique is that it has no effect on the traffic being monitored. Passive monitoring can only be detected by launching a probe-response attack (Bethencourt *et al.*, 2005) on the monitoring system. When elements of intrusion detection systems are "probed" by intentionally suspicious traffic they generate an alert that can be detected by a third party and thus reveal the location of these elements. Similarly, a communication that is seemingly of interest to the eavesdroppers can be routed through some nodes of the network. If the eavesdroppers act after seeing this communication it means that these nodes are under surveillance. Unfortunately, the actions of eavesdroppers are not as well-defined as the

operation of intrusion detection systems and may not be seen by the third party trying to detect their activities. Nevertheless, if eavesdropped nodes are detected, sophisticated Internet users may route their communications to avoid the eavesdropped nodes and so preserve their privacy.

Naïve Internet users may assume that since the Internet network is so large, their communication tracks will blend into the communication tracks of others. In the rest of the paper, we show that such an assumption is quite misleading even in the presence of above-mentioned governmental protection. If several eavesdroppers collaborate in order to compromise Internet users' anonymity and choose the nodes to be monitored in an optimal manner, there is a significant chance that they can track down the communications of many individual users.

2.2. Complex topologies and BC measure

A variety of studies aimed at comprehending the generic features of complex network development have shown that many systems display topologies that often seem random and unpredictable. Two examples of complex networks are in the social sciences, and the Internet. In the social sciences nodes are persons or organizations while the edges characterize the relations between them (Wasserman and Faust, 1994). In the Internet, nodes are routers/autonomous-systems and edges are the links connecting one router/autonomous-system to another.

Surprisingly, research has shown that complex networks self-organize into a scale-free structure and that their node connectivity distribution follows a power-law (Barabási and Albert, 1999; Barabási *et al.*, 2000). Therefore, independent of the nature of the system and the identity of its constituents, the probability $P(k)$ that a node in the network is connected to k other nodes decays as a power-law, following $P(k) \sim k^{-\gamma}$. The Internet, for example, is a growing network interconnecting millions of computers around the world. It has been shown that despite the Internet's apparent random character, its topology has a universal scale-free characteristic that is determined by the web's connectivity (Barabási *et al.*, 2000; Vazquez *et al.*, 2002). This characteristic is found to be a consequence of two generic mechanisms: the continuous expansion of networks due to additions of new nodes and the preferential attachment of new nodes to already well-connected nodes (Barabási and Albert, 1999).

In order to analyze and understand the roles played by nodes in complex networks, many network-analytic studies in recent years have relied on the evaluation of centrality measures defined for the nodes of the network. These measures have been used to rank the individual node's prominence according to its position in the network (Wasserman and Faust, 1994).

An important centrality measure that we will be focusing on in the current study is the shortest path BC defined by Freeman (1977). BC measures the extent to which a node lies on the paths between other nodes. The BC of a node is defined as the sum of fractions of shortest paths between pairs of nodes in a network in which the node takes part. High BC scores indicate a high ratio of shortest paths that a node lies on. Various modifications of the original BC measure were summarized by Brandes (2008), along with efficient algorithms for their computation.

BC was designed as a measure of the extent to which a node has control over information flow in communication networks. It is even used by Yan *et al.* (2006) to

administer a routing scheme that avoids potential congestions. BC can also be used by an eavesdropper seeking to launch an attack on anonymity on the Internet. The purpose of this attack on anonymity is to unveil the identity of communicating nodes. The identity of the communicating nodes can be traced by eavesdropping on one node on the path along which the information flows between two of them. By applying the centrality measure, an eavesdropper can measure how much information will pass through each node in the network, and eavesdrop on nodes that are more likely to contribute to his/her objective. The more pairs of nodes that communicate through an eavesdropped node, the greater the probability the attack on anonymity will succeed.

Furthermore, by strategically choosing a group of collaborating nodes to eavesdrop on, an eavesdropper may gain more control over information flow. An eavesdropper cannot rely on the centrality measure of a singe node while trying to calculate the effectiveness of a group of collaborating nodes. This is due to the redundant contribution certain nodes may have while collaborating with other nodes, for example by covering the same communication paths. Instead, the attacker can try to maximize the number of distinct communication paths covered by the group.

3. Locating eavesdroppers on nodes with high centrality

Next we formalize the methods that can be exploited by attackers to locate eavesdroppers on key positions in the Internet topology. Although the terminology used in this section is fairly intuitive, the reader is suggested to look into West (2001) and Brandes (2008) for an introduction into graph theory and BC. Let us assume an eavesdropper is seeking to launch an attack on the anonymity of Internet users. Let $G = (V, E)$ describe the Internet network, where $v \in V$ represent the nodes in the network, and $e \in E$ represent the links between the nodes. Let n describe the number of nodes and m describe the number of links. We further assume that all graphs presented in our study are undirected, connected and un-weighted.

The eavesdropper's objective is to unveil the identity of communicating parties. The eavesdropper may achieve his objective by eavesdropping on nodes that lie on as many communication paths as possible. At this preliminary stage, we assume that our eavesdropper has the resources to eavesdrop on one single node. This assumption requires the eavesdropper to carefully choose the node through which most information flows. By applying the BC measure, an eavesdropper can estimate how much information will pass through each node in the network and locate nodes that are more likely to contribute to its objective.

Standard measures of BC can be used to measure a node's importance in a graph. High BC scores indicate that a node lies on many fractions of shortest paths connecting others (Freeman, 1977), therefore eavesdropping such nodes increases the success rate of an attack on anonymity. The shortest path from node s to node t is the path we get by moving from s to its predecessors, and then to the predecessors of each successive node, until t is reached. If a node has two or more predecessors, then there may be more than one shortest path between s and t, which may or may not share some nodes. Let σ_{st} denote the number of shortest paths from node s to node t. Let $\sigma_{st}(v)$ denote the number of shortest paths from node s to node t on which some node v lies. For example, in the network given in Figure 1, it can be seen that there are four shortest paths between node 1 and node 10: {1-2-3-5-8-10}, {1-2-3-6-8-10}, {1-2-3-6-9-10} and

{1-2-4-7-9-10}. Therefore, given $s = 1$ and $t = 10$, $\sigma_{st} = 4$. Furthermore, given $v = 6$, $\sigma_{st}(6) = 2$, since there are only two shortest paths from s to t on which v lies.

Brandes (2001) denotes $\delta_{st}(v) = \sigma_{st}(v)/\sigma_{st}$ as the pair-dependency of a pair s, t on node v (a node that lies on a shortest path between s and t). Following our example based on Figure 1, where $s = 1$, $t = 10$ and $v = 6$: $\delta_{st}(v) = \sigma_{st}(v)/\sigma_{st} = 2/4 = 1/2$. Brandes denotes $\delta_{s*}(v)$ as the dependency of the communications emanating from s on a single node v:

$$\delta_{s*}(v) = \sum_{\substack{t \in V \\ s \neq t \neq v}} \delta_{st}(v) = \sum_{\substack{t \in V \\ s \neq t \neq v}} \frac{\sigma_{st}(v)}{\sigma_{st}}.$$

Let $C_B(v)$ denote the BC for node v:

$$C_B(v) = \sum\nolimits_{s \in V} \delta_{s*}(v).$$

Algebraic path counting can be used to compute BC of all vertices in the graph in $O(n^3)$. Brandes (2001) suggested a more efficient algorithm that eliminates the need for summation of all pair-dependencies. While iterating over all source vertices in the graph it calculates the dependency of the source on all other nodes. At the end of the iteration, the dependency calculated for each node is added to the BC score of that node.

Assume every pair of nodes generates the same communication flow in the network. Let $S(v)$ denote the percentage of communicating pairs of nodes an eavesdropper can monitor at v. $S(v)$ can be calculated by dividing the $C_B(v)$ by the total number of communicating pairs C_n^2 where n is the number of nodes in the network:

$$S(v) = C_B(v)/C_n^2.$$

Accordingly, we define the measure of global anonymity of all users in the network when the eavesdropper is monitoring a single node v ($A(v)$) by the following equation:

$$A(v) = 1 - C_B(v)/C_n^2.$$

A higher BC of a node that is monitored by an eavesdropper results in lower anonymity of the users in the network. $A(v) = 0$ means that eavesdroppers have the

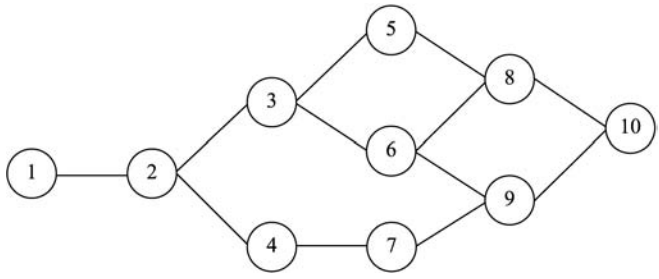

Figure 1.
Example of a connected, undirected and un-weighted network with ten nodes where $n = 10$ and $m = 12$

ability to monitor all the communication in the network and all the communicating nodes' anonymity is compromised.

4. Group of collaborating eavesdroppers

As demonstrated in the previous section, an eavesdropper devising an attack on anonymity may enjoy a higher success rate by eavesdropping on a node with higher BC. In many scenarios, eavesdropping on a single node may not be sufficient to accomplish a successful attack on anonymity. In such a case, a collaborative attack of a group of nodes that are eavesdropped may prove to be more efficient. The intuitive way to use the BC measure while devising a collaborative attack would be to sort the nodes by their BC values and eavesdrop on a group of nodes with the highest BC values. This basic intuitive approach is not free of doubts. The contribution of a node's BC to a group cannot be added linearly. This limitation stems from the nature of the BC value, which does not take into account the role of other nodes in the group. For example, a pair of attackers may be positioned on two neighboring nodes, each having a very high BC value. Nonetheless, these two nodes may be located on the same shortest paths so that neither of the attackers contributes novel information to the other.

In order to define the success of eavesdropping on a group of nodes we use group betweenness centrality (GBC), as defined by Everett and Borgatti (1999). GBC of a group of nodes stands for the fraction of shortest paths between pairs of nodes in a network that passes through at least one of the nodes in the group. Thus, GBC can be used for estimating the ability of a group of collaborating nodes to monitor the network traffic.

GBC can be viewed from an additional utilitarian point of view. While integrating the group centrality calculations as part of a comprehensive strategy, an eavesdropper can exploit the GBC measure in order to minimize the overall costs, while acquiring sufficient malicious nodes across the network. The basic assumption underlying this conclusion is that an attacker seeks to gain as much information as possible with minimum effort or cost. For example, an eavesdropper seeking to launch an attack on anonymity on the Internet may gather as much information as possible in order to obtain the identity of communicating individuals. Obtaining such information is possible by eavesdropping on nodes located on the communication paths of as many communicating pairs as possible. The GBC measure is suggested for the purpose of addressing this need.

We have modified the algorithm for computing the BC of all vertices (Brandes, 2001) to compute the GBC of one group of nodes in the same asymptotic running time (Yagil, 2005) (see the Appendix for explanations of the algorithm along with complete pseudo-code).

Algorithm 1: Computing the GBC of a group
Input. The graph $G = (V, E)$ and the group $(v1, \ldots, vk) \subseteq V$.
Output. GBC (initialized to zero).

Calculation

(1) Loop through all nodes $s \in V$:

- Let S be an empty stack that will be used to sort nodes in the order of non-increasing distance form s.
- Let $P_{s,w}$ be a list of predecessors of w on the way from s.
- Let $d_{s,t}$ be the distance between nodes s and t. $d_{s,t}$ is initalized to -1 for each s and t except $d_{s,s}$ which is equal to zero.
- Let $\sigma_{s,t}$ be the number of shortest paths between s and t. $\sigma_{s,t}$ is initalized to zero except $\sigma_{s,t}$ which is equal to 1.
- Let Q be a queue iniitaly containing only s. Q will be used to traverse G in breadth first order.

(2) While Q is not empty:
- move the node v from Q to S;
- for each neighbor w of v encountered for the first time ($d_{s,w} = -1$) add w to Q and update $d_{s,w}$ to $d_{s,v} + 1$; and
- for each neighbor of w of v such that v lays on a shortest path from s to w ($d_{s,w} = d_{s,v} + 1$) add v to $P_{s,w}$ and update $\sigma_{s,w}$ to $\sigma_{s,w} + \sigma_{s,v}$.

(3) Let $\delta'_{s,*}(v)$ be the influence of v on communications emanating from s. $\delta'_{s,*}(v)$ is initialized to zero.

(4) Traverse the nodes in the graph in the order of non-increasing distance from s by popping each next node (w) from S.

(5) For each predecessor ($v \in P_{s,w}$) of w add $\sigma_{s,v}/\sigma_{s,w}*(1 + \delta'_{s,*}(w))$to $\delta'_{s,*}(v)$ if w is not one of the input vertices $\{v1, \ldots vk\}$, otherwise, add $\delta'_{s,*}(w)$ to GBC and set $\delta'_{s,*}(v)$ to zero.

The same algorithm was independently proposed by Brandes (2008). A different algorithm that efficiently computes the GBC of many groups on the same network was proposed by Puzis *et al.* (2007b).

Let $S(v_1, \ldots, v_k)$ denote the success of an eavesdropper in intercepting communications between nodes, while monitoring all the communication of nodes$v_1, .., v_k$. $S(v_1, \ldots, v_k)$ can be calculated by dividing the GBC$(v_1, \ldots, v_k)$ by the total number of communicating pairsC_n^2, where n is the number of nodes in the network. The greater the GBC of a group of nodes is, the greater the success of the eavesdropper located on these nodes:

$$S(v_1, \ldots, v_k) = \mathrm{GBC}(v_1, \ldots, v_k)/C_n^2.$$

Accordingly, the measure of global anonymity of all the users in the network, when the eavesdropper is monitoring nodes $v_1, .., v_k$ (denoted by$A(v_1, \ldots, v_k)$), may be defined by the following equation:

$$A(v_1, \ldots, v_k) = 1 - \mathrm{GBC}(v_1, \ldots, v_k)/C_n^2.$$

Higher GBC values of a group of nodes that are monitored by an eavesdropper signify lower anonymity levels of the users of the network. An $A(v_1, \ldots, v_k)$ value of one would indicate that there are no eavesdroppers that are capable of linking two

communicating parties. On the other hand, a value of zero would indicate that any packet sent over the Internet can be intercepted by collaborating eavesdroppers. Roughly, $A(v_1, \ldots, v_k)$ indicates the chance that a communication of two random parties cannot be intercepted.

It is reasonable to assume that the effort required to eavesdrop a group of nodes ($F(v_1, \ldots, v_k)$) is proportional to the total amount of analyzed traffic. BC is highly correlated with the amount of traffic passing through a node in communication networks (Holme, 2003; Yan *et al.*, 2006). Therefore, we use the sum of the BC of the group members as a measure of the attackers' effort:

$$F(v_1, \ldots, v_k) = \sum_{i=1}^{k} \mathrm{BC}(v_1, \ldots, v_k)/C_n^2.$$

5. Evaluating an attack on anonymity on the Internet

In this section we evaluate the ease or difficulty of compromising the anonymity of Internet users by a group of collaborating eavesdroppers when the most central nodes are protected. In order to perform the evaluations, we used a snapshot of the Internet autonomous systems level topology from March 24, 2008 made available by CAIDA. We executed the simulations on an un-weighted undirected version of the network. The network contained 21,823 nodes and 89,866 links, with an average degree of 4.11 and power-law exponent of degree distribution of -1.12. Only 2250 ($\sim$1 percent) of the nodes in the network have degrees above the average. Due to the scale-free structure of the investigated network a few most central nodes cover almost all the Internet traffic. We assumed that the attackers were eavesdropping on 1 to 100 nodes from the 5000 ($\sim$23 percent) most central nodes in the network, while 1 to 18 of the most central nodes were protected from eavesdropping. We have tested two attacker deployment strategies. In the first strategy (TopK) the attackers chose the nodes with highest BC (K nodes with the highest). In the second strategy (GreedyKGBC) the attackers constructed the group of nodes iteratively by choosing each time a node with greatest contribution to the GBC of the already chosen nodes (Puzis *et al.*, 2007a). The simulation results are presented in Figures 2-4.

The anonymity level as the function of the size of the group of attackers is presented in Figure 2. The deployment strategy here was GreedyKGBC, as it produced results better than or equal to TopK in all simulations. The different series in Figure 2 correspond to the number of protected nodes. For example, if the ten most central nodes are protected from eavesdropping it is enough to eavesdrop on 48 less central nodes to reduce the anonymity level to 0.4 (see Figure 2 left). In this paper we derive the anonymity level from the GBC value of the group of eavesdropped nodes and the effort required to eavesdrop on the group of nodes from the sum of BC values of the respective nodes. The relationship between the anonymity level and the eavesdropping effort of groups found by TopK and GreedyKGBC strategies are shown in Figure 2 (right). In simulations summarized in this figure, nodes were not protected from eavesdropping. We can see that when there are few collaborating nodes the attackers gain roughly what they are paying for. However, when network coverage increases, the effort required to increase it even more is enormously high compared to the potential

gain. The reason is redundant inspection of the same traffic by many nodes in the network. We can also see that when network coverage increases, the difference between two deployment strategies becomes apparent. This is more pronounced in Figure 4.

We can see in Figure 3 (left), that when some of the most central nodes are protected the attacker needs to eavesdrop more nodes in order to compromise anonymity in the

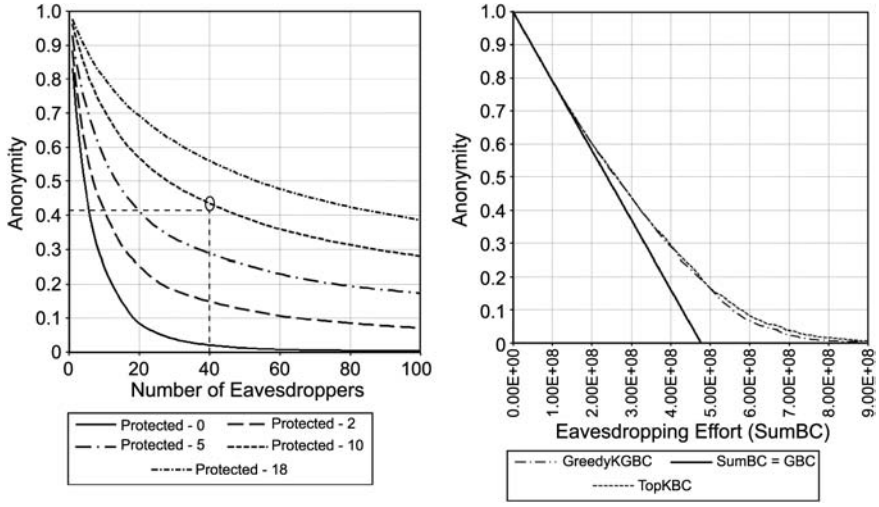

Figure 2.
Anonymity level as a function of the number of collaborating eavesdroppers chosen by the GreedyKGBC algorithm (left); anonymity level as a function of the eavesdropping effort without protected nodes (right)

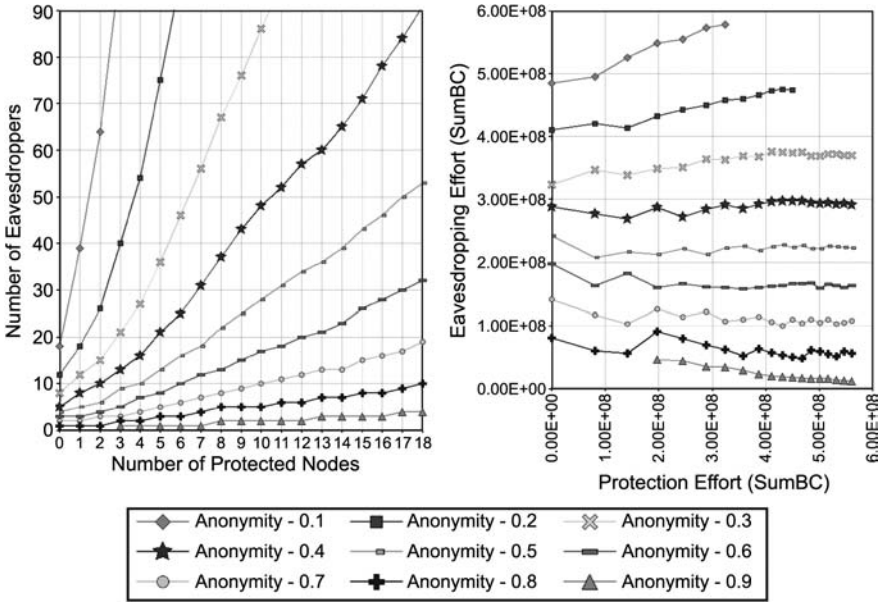

Figure 3.
Eavesdropping cost vs protection cost (left – in terms of the number of nodes; right – in terms of SumBC)

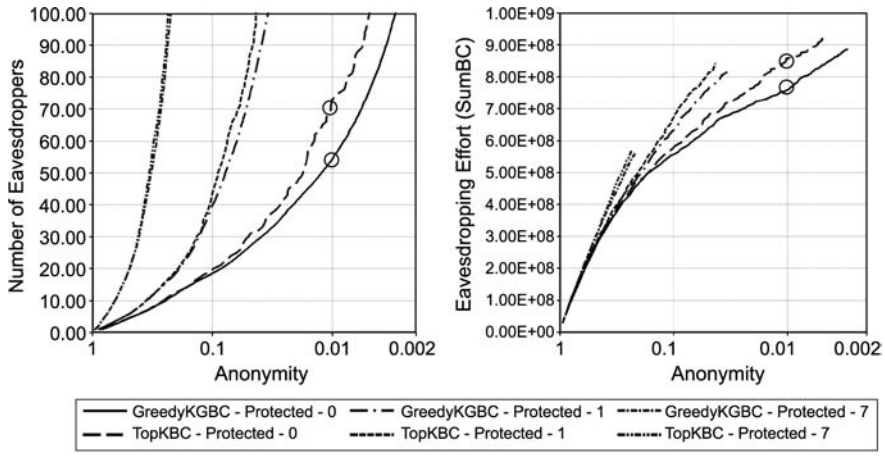

Figure 4.
Comparison of two eavesdroppers' deployment strategies

network with the same level of success. However, bandwidth requirements remain roughly the same as can be noticed from Figure 3 (right). We can also see in Figure 3 (left) that when the attackers' target is to reduce the anonymity level to 0.7 the number of eavesdropped nodes is roughly equal to the number of protected nodes. Moreover, protected nodes have a higher individual BC than the eavesdropped nodes. If we assume that the attackers' eavesdropping effort is equal to the effort required to protect the nodes from eavesdropping, then Figure 3 (right) suggests that such protection is highly inefficient.

Finally we would like to compare the two strategies for deployment of collaborating eavesdroppers. We can see in Figure 4 that when the attacker wants to compromise the anonymity of most of the Internet users (anonymity level lower than 0.05), it is better to use GreedyKGBC to locate the optimal group despite the required computational effort. For example, in order to reduce the anonymity level to 0.01 the attackers may eavesdrop on 54 nodes instead of the 73 nodes suggested by the TopK strategy, reducing the number of eavesdropped nodes by approximately 25 percent (see Figure 4 left). Similarly, GreedyKGBC reduces the eavesdropping effort by 8.9 percent compared to TopKBC (see Figure 4 right). Our simulations show that the difference between the two strategies decreases as more most-central-nodes are protected.

6. Conclusion and future work

In this study we analyzed an attack by a group of eavesdroppers who collaborate to compromise the Internet users' anonymity by eavesdropping on several nodes on the Internet. We assumed that the Internet users under attack are not using any anonymity enhancing technologies and that the most prominent nodes may be protected from eavesdropping. Possible future research not covered by this paper includes analysis of the global anonymity level in the case where the most prominent nodes are not only protected from eavesdropping but also supply anonymity services. Future research may also include a more precise cost model or investigating anonymity of Internet communities in the presence of global eavesdroppers.

We have demonstrated how GBC can be used to evaluate the level of anonymity, given a group of collaborating eavesdroppers on the Internet. We illustrated that it is possible to compromise the anonymity of many Internet users when eavesdropping on a relatively small number of nodes. We also show that these nodes are not necessarily the most prominent nodes. Our study demonstrates the ease of a significant attack on anonymity even when the most prominent nodes in the network cannot be eavesdropped. According to assumptions and findings of this work it is not cost efficient to prevent autonomous systems from being eavesdropped. A good alternative would be regulations that will prevent eavesdropped autonomous systems from collaborating with one another.

This study also emphasizes that when trying to completely compromise the anonymity of Internet users, an eavesdropper's deployment method based on GBC can result in substantial resource saving compared to choosing a set of the most prominent nodes.

To summarize, this study complements previous studies in the field of Internet topology and adds to our understanding of interactions between nodes within a network.

References

Barabási, A.-L. and Albert, R. (1999), "Emergence of scaling in random networks", *Science*, No. 286, pp. 509-12, available at: http://arXiv.org/abs/cond-mat/9910332

Barabási, A.-L., Albert, R. and Jeong, H. (2000), "Scale-free characteristics of random networks: the topology of the world-wide web", *Physica A*, No. 281, pp. 69-77.

Barthélemy, M. (2004), "Betweenness centrality in large complex networks", *European Physical Journal B*, Vol. 38 No. 2, pp. 163-8.

Berthold, O., Federrath, H. and Köhntopp, M. (2000), "Project anonymity and unobservability in the Internet", *Proceedings of the Tenth Conference on Computers, Freedom and Privacy: Challenging the Assumptions*, pp. 57-68.

Bethencourt, J., Franklin, J. and Vernon, M. (2005), "Mapping internet sensors with probe response attacks", *Proceedings of the 14th Conference on USENIX Security Symposium, Berkeley, CA*, Vol. 14, p. 13.

Boyan, J. (1997), "The anonymizer – protecting user privacy on the web", *Computer-mediated Communication Magazine*, Vol. 4 No. 9, pp. 7-13, available at: www.december.com/cmc/mag/1997/sep/boyan.html

Brandes, U. (2001), "A faster algorithm for betweenness centrality", *Journal of Mathematical Sociology*, Vol. 25 No. 2, pp. 163-77.

Brandes, U. (2008), "On variants of shortest-path betweenness centrality and their generic computation", *Social Networks*, Vol. 30 No. 2, pp. 136-45.

Cai, M., Hwang, K., Kwok, Y., Song, S. and Chen, Y. (2005), "Collaborative Internet worm containment", *IEEE Security and Privacy*, Vol. 3 No. 3, pp. 25-33.

CAIDA (2008), "AS relationships dataset", March 24, avilable at: www.caida.org/data/active/as-relationships/ (accessed October 21, 2008).

Chaum, D. (1981), "Untraceable electronic mail, return addresses, and digital pseudonyms", *Communications of the ACM*, Vol. 24 No. 2, pp. 84-8.

Claessens, J., Preneel, B. and Vandewalle, J. (1999), "Solutions for anonymous communication on the internet", *Proceedings of the 1999 IEEE International Carnahan Conference on Security Technology*, pp. 298-303.

Claessens, J., Díaz, C., Goemans, C., Dumortier, J., Preneel, B. and Vandewalle, J. (2003), "Revocable anonymous access to the Internet?", *Internet Research*, Vol. 13 No. 4, pp. 242-58.

Díaz, C., Seys, S., Claessens, J. and Preneel, B. (2003), "Towards measuring anonymity", in Dingledine, R. and Syverson, P. (Eds), *Designing Privacy Enhancing Technologies*, LNCS 2009, Springer-Verlag, Berlin and Heidelberg, pp. 54-68.

Edman, M., Sivrikaya, F. and Yener, B. (2007), "A combinatorial approach to measuring anonymity", *Intelligence and Security Informatics, 2007*, IEEE, New York, NY, pp. 356-63.

Elovici, Y. and Glezer, C. (2003), "Anonymity enhancing technologies (AET): opportunities and threats", *The Journal of Information Warfare*, Vol. 3 No. 3, pp. 48-64.

Everett, M.G. and Borgatti, S. (1999), "The centrality of groups and classes", *Mathematical Sociology*, Vol. 23 No. 3, pp. 181-201.

Faloutsos, M., Faloutsos, P. and Faloutsos, C. (1999), "On power-law relationships of the Internet topology", *Computer Communication Review*, Vol. 29 No. 4, pp. 251-62.

Freedman, M.J. and Morris, R. (2002), "Tarzan: a peer-to-peer anonymizing network layer", *Proceedings of the ACM Conference on Computer and Communications Security (CCS 9)*.

Freeman, L.C. (1977), "A set of measuring centrality based on betweenness", *Sociometry*, Vol. 40 No. 1, pp. 35-41.

Gabber, E., Gibbons, P.B., Matias, Y. and Mayer, A. (1997), "How to make personalized web browsing simple, secure, and anonymous", *Proceedings of Financial Cryptography*, LNCS 1318, Springer-Verlag, Berlin, pp. 17-31.

Goldschlag, D.M., Reed, M.G. and Syverson, P.F. (1999), "Onion routing for anonymous and private Internet connections", *Communications of the ACM*, Vol. 42 No. 2, pp. 39-41.

Gritzalis, S. (2004), "Enhancing web privacy and anonymity in the digital era", *Information Management & Computer Security*, Vol. 12 No. 3, pp. 255-88.

Holme, P. (2003), "Congestion and centrality in traffic flow on complex networks", *Advances in Complex Systems*, Vol. 6 No. 2, p. 163.

Pfitzmann, A. and Hansen, M. (2008), "Anonymity, unlinkability, unobservability and pseudonymity, and identity management – a consolidated proposal for terminology v0.31", February, avilable at: http://dud.inf.tu-dresden.de/literatur

Pfitzmann, A. and Waidner, M. (1986), "Networks without user observability – design options", *Eurocrypt 85*, LNCS 219, Vol. 219, Springer-Verlag, Berlin, pp. 245-53.

Pfitzmann, A. and Waidner, M. (1989), "Networks without user observability", *Computer & Security*, Vol. 2 No. 6, pp. 158-66.

Puzis, R., Elovici, Y. and Dolev, S. (2007a), "Finding the most prominent group in complex networks", *AI Communications*, Vol. 20 No. 4, pp. 287-96.

Puzis, R., Elovici, Y. and Dolev, S. (2007b), "Fast algorithm for successive computation of Group Betweenness Centrality", *Physical Review E*, Vol. 76 No. 5, p. 056709.

Raymond, J.F. (2001), "Traffic analysis: protocols, attacks, design issues and open problems", in Federrath, H. (Ed.), *Proceedings of International Workshop on Design Issues in Anonymity and Unobservability*, LNCS 2009, Springer-Verlag, Berlin, pp. 10-29.

Reiter, M.K. and Rubin, A.D. (1998), "Crowds: anonymity for web transactions", *ACM Transactions on Information and System Security*, Vol. 1 No. 1, pp. 66-92.

Reiter, M.K. and Rubin, A.D. (1999), "Anonymous web transactions with crowds", *Communications of the ACM*, Vol. 42 No. 2, pp. 66-92.

Serjantov, A. and Danezis, G. (2002), "Towards an information theoretic metric for anonymity", in Dingledine, R. and Syverson, P. (Eds), *Designing Privacy Enhancing Technologies*, LNCS 2009, Springer-Verlag, Berlin and Heidelberg, pp. 41-53.

Shapira, B., Elovici, Y., Mashiach, A. and Kuplik, T. (2005), "PRAW – a new model for PRivAte Web navigation", *Journal of the American Society of Information Science Technology*, Vol. 56 No. 2, pp. 159-72.

Shields, C. and Levine, B.N. (2000), "A protocol for anonymous communication over the internet", *ACM Conference on Computer and Communications Security*, pp. 33-42.

Shields, C. and Levine, B.N. (2002), "Hordes: a protocol for anonymous communication over the Internet", *ACM Journal of Computer Security*, Vol. 10 No. 3, pp. 213-40.

Shmatikov, V. and Wang, M. (2007), "Security against probe-response attacks in collaborative intrusion detection", *Proceedings of the 2007 Workshop on Large Scale Attack Defense, LSAD 2007*, Vol. 2007, p. 2007.

Song, R. and Korba, L. (2002), "Review of network-based approaches for privacy", *Proceedings of the 14th Annual Canadian Information Technology Security Symposium, Ottawa*, pp. 1-10.

Strogatz, S.H. (2001), "Exploring complex networks", *Nature*, No. 410, pp. 268-76.

Vazquez, A., Pastor-Satorras, R. and Vespignani, A. (2002), "Large-scale topological and dynamical properties of Internet", *Physical Review E*, Vol. 65 No. 066130, available at: http://arXiv.org/abs/cond-mat/0112400

Wasserman, S. and Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge.

West, D.B. (2001), *Introduction to Graph Theory – Second Edition*, Prentice-Hall, Upper Saddle River, NJ.

Yagil, D. (2005), "Collaborative attack on WWW users' anonymity", master's thesis, Department of Information Systems Engineering, Ben-Gurion University, Beer Sheva.

Yan, G., Zhou, T., Hu, B., Fu, Z.-Q. and Wang, B.-H. (2006), "Efficient routing on complex networks", *Physical Review E*, Vol. 73 No. 046108.

Yegneswaran, V., Barford, P. and Jha, S. (2004), "Global intrusion detection in the DOMINO overlay system", *Proceedings of Network and Distributed System Security Symposium NDSS 2004*.

Appendix. Pseudo-code of the algorithm for GBC computation

The pseudo code is shown in Figure A1.

The input of algorithm 1 is the graph $G = (V, E)$ and a group of nodes $v_1, .., v_k$. Algorithm 1 computes GBC of $v_1, .., v_k$. The algorithm iterates over all the nodes in V (line 2). Let s be the current source node. In lines 3-27 of the algorithm for each node $w \in V$, it computes the distance from s to w ($d_{s,w}$), the number of shortest paths between s and w ($\sigma_{s,w}$), and a set of predecessors leading from w to s ($P_{s,w}$). In addition, all nodes on the graph are ordered in non-increasing distance from s. Let $\delta_{s,*}(v_1, \ldots, v_k)$ denote the dependency of s on the group$v_1, .., v_k$.

Let $\delta'_{s,*}(v_i)$ be the contribution of node v_i to $\delta_{s,*}(v_1, \ldots, v_k)$:

$$\delta_{s,*}(v_1, \ldots, v_k) = \sum_{i \in \{1, \ldots, k\}} \delta'_{s,*}(v_i)$$

```
GBC(v1,…,vk):
    Input: G=(V,E), v1…vk
    Output: GBC(v1…vk)
    Calculation:
1.  C'_B(v) ← 0, v ∈ V;
2.  for s ∈ V do:
3.          B ← v_1,...,v_k;
4.          A ← empty list;
5.          S ← empty stack;
6.          P_{s,w} ← empty list, w ∈ V;
7.          σ_{s,t} ← 0, t ∈ V, σ_{s,s} ← 1;
8.          d_{s,t} ← −1, t ∈ V, d_{s,s} ← 0;
9.          Q ← empty queue;
10.         enqueue s → Q;
11.         //calculating # of shortest paths with BFS algorithm from s to all other nodes
12.         while Q not empty do
13.                 dequeue v ← Q;
14.                 push v → S;
15.                 for each neighbor w of v do
16.                         // w found for the first time?
17.                         if d_{s,w} < 0 then
18.                                 enqueue w → Q;
19.                                 d_{s,w} ← d_{s,v} + 1;
20.                         end
21.                         // shortest path to w via v?
22.                         if d_{s,w} = d_{s,v} + 1 then
23.                                 σ_{s,w} ← σ_{s,w} + σ_{s,v};
24.                                 append v → P_{s,w};
25.                         end
26.                 end
27.         end
28.         δ'_{s,*}(v) ← 0, v ∈ V;
29.         // S returns nodes in order of non-increasing distance from s
30.         A ← B;
31.         while S not empty and A not empty do
32.                 pop w ← S;
33.                 if w ∈ A then A = A − w;
34.                 for v ∈ P_{s,w} do
35.                         if w ∉ B then
36.                                 δ'_{s,*}(v) ← δ'_{s,*}(v) + (σ_{s,v} / σ_{s,w}) * (1 + δ'_{s,*}(w));
37.                         else
38.                                 δ'_{s,*}(v) ← 0;
39.                         end
40.                 end
41.                 if w ≠ s then C'_B(w) ← C'_B(w) + δ'_{s,*}(w);
42.         end
43. end
44. GBC(v_1,…,v_k) = ∑ C_B[v_i]
```

Figure A1.

The second part of the algorithm, lines 28-42, iterates over all nodes in the graph in order of non-increasing distance from s, computing $\delta'_{s,*}(v_i)$ of all group members.

Let w be the node chosen in line 32. In line 36 $\delta'_{s,*}(w)$ considers shortest paths that start at s, pass through w, and do not pass through members of the group $v_1, \ldots, v_k$ such that $v \neq w$ and $d_{s,v_i} \geq d_{s,w}.d'_{s,*}(v)$ is accumulated along the shortest path to s (line 36) until we reach a group member. Group members do not contribute to dependency of s on their predecessors since we want to consider each shortest path only once.

Line 41 accumulates the dependency of s on w in order to get the contribution of w to GBC($v_1, \ldots, v_k$) considering all source vertices. Line 44 of the algorithm sums the contributions of all the group members $v_1, \ldots, v_k$ to GBC($v_1, \ldots, v_k$).

Corresponding author
Rami Puzis can be contacted at: puzis@bgu.ac.il