Group-Based Characterisation for the I2P Anonymous File-Sharing Environment

Juan Pablo Timpanaro, Isabelle Chrisment, Olivier Festor

Group-Based Characterisation for the I2P Anonymous File-Sharing Environment

Juan Pablo Timpanaro†, Isabelle Chrisment*, Olivier Festor†
†INRIA, Villers-lès-Nancy, F-54600, France
*Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-les-Nancy, F-54506, France
Email: {juanpablo.timpanaro, olivier.festor}@inria.fr
Email: {isabelle.chrisment}@loria.fr

***Abstract*—The I2P network provides an abstraction layer allowing two parties to communicate in an anonymous manner. This network is optimised for anonymous web hosting and anonymous file-sharing. I2P's file-sharing community is highly active where users deploy their file-sharing applications on top of the network. I2P uses a variation of Onion routing, thus assuring the unlinkability between a user and its file-sharing application.**

In this paper, we take the first step towards the linkability of users and applications in the I2P network. We conduct a *group-based characterisation*, where we determine to what extent a group of users is responsible for the overall I2P's file-sharing activity. We used Pearson's coefficient to correlate users from two cities and the most used anonymous file-sharing application. We determine that two cities explain more than a third of all file-sharing activity within the I2P network.

***Index Terms*—Group-based characterisation; Anonymous networks; Linkability; I2P; Pearson's correlation**

I. Introduction

Anonymous communications are growing fast[1] because they allow users to access different services while preserving their online privacy. These systems conceal a user's real identity by usually decoupling it from the assigned system's identity. Therefore, by supervising these systems we are able to characterise users and applications independently, but we cannot link the behaviour of both and determine to what extent they are related.

Our goal is to show that it is possible to infer the implication of a group of users for a given application's activity, even in an anonymous environment such as the I2P network. Our contribution is the first group-based characterisation, where we target the most active environment in the I2P network, its BitTorrent-like file-sharing environment. In our approach, we monitor at the same time a specific application and users' behaviour to conduct a comprehensive correlation analysis based on data collected from our distributed monitoring architecture [1]. The Pearson's coefficient is used to correlate the top detected cities with the most used I2P file-sharing application, to determine whether these cities explain, and in which measure, to I2P's file-sharing activity.

[1]The Tor network has tripled its user-base in the last three years, while the I2P anonymous network has doubled its user-base in the last year. Statistics from //metrics.torproject.org and //stats.i2p.in/, respectively. Last visited on 10/2013.

The rest of the paper is organised as follows: Section 2 introduces the I2P network. Section 3 describes our correlation strategy and details Pearson's coefficient. Section 4 presents our experimental results. Section 5 discusses the privacy implications of our approach, as well as the ethic aspects of monitoring an anonymous system. Section 6 gives a background on monitoring analyses related to anonymous systems. Finally, Section 7 concludes our work.

II. The I2P Network

The *Invisible Internet Project*, or also known as I2P, is mainly designed for anonymous web hosting and anonymous file-sharing. Except for anonymous web browsing that necessarily requires an out-proxy to the Internet, the rest of the applications interact among each other within the network boundaries. The system is designed as an anonymous network layer, enabling users to deploy their own applications on top of the network. On the contrary to the Tor network [2], where users' traffic enters the network, gets re-routed and exits to the normal Internet, within the I2P network the traffic stays on the network. Here we detail the main characteristics of this network.

A. I2P's anonymity

The I2P network layer is composed of participants known as I2P *nodes* or I2P *routers*, where every node in the system forwards traffic on behalf of the network. Information regarding every particular I2P router, *e.g.* its IP address, is gathered in a special structure called *routerinfo*. An I2P node uses *tunnels* to communicate with other nodes, where these tunnels are formed by others I2P nodes. Whenever an I2P router `A` wants to create a tunnel with an I2P router `B`, router `A` needs to get router `B`'s routerinfo.

In the I2P network an user's application is not identified through the tuple <IP address, port number>, but via a location-independent identifier which decouples a user's online identity and her/his actual physical location. This hash-like identifier is known as a *destination*. Every time a user deploys an I2P application, such as a file-sharing client, a destination is created for that application. This destination is used to receive incoming messages from third-parties, such as other I2P file-sharing clients. Information concerning a particular destination is grouped in a special structure called *leaseset*. A remote

I2P user needs this leaseset to establish a connection with the application.

The basis of I2P's anonymity is the unlinkability between leasesets and routerinfos. It is not possible to link a particular leaseset, representing an application, with a particular routerinfo, representing an I2P user. Let's illustrate this scenario with two I2P users, running two BitTorrent applications on top of the I2P network. Each user, `A` and `B`, has its own routerinfo ri_a and ri_b, respectively. Each BitTorrent application, app_a and app_b, has its own leaseset ls_a and ls_b, respectively. In the I2P network, it is not possible to link ri_a with ls_a or ri_b with ls_b and therefore determine that the user `A` is running the application app_a, for instance.

A leaseset has a set of entry points or *gateways* to receive messages from third-parties. Ls_a will have one (or more) entry points, where remote applications, *e.g.* app_b, can send messages. These entry points are the I2P nodes in the end of the tunnel of the user `A`, which are represented with different routerinfos. Therefore, an application app_a will have a leaseset ls_a, where the entry points are ri_x and ri_y. The remote application app_b will communicate with app_a through ri_x and ri_y.

B. Distributed network database

I2P uses a distributed database to store its *network metadata*, that is, leasesets and routerinfos. The database is called the *netDB* and is a Kademlia-based [3] distributed hash table, composed of *floodfill* nodes. Floodfill nodes are normal I2P nodes with high bandwidth rates. All routerinfos and leasesets are stored within these floodfill nodes, and are accessible by every node in the I2P network.

Considering the previous example, ri_a, ri_b, ls_a and ls_b are stored within the netDB. The I2P user `A` running the application app_a has a destination $dest_a$ and its associated leaseset ls_a. If the I2P users `B` running the application app_b wants to contact the application app_a, it needs to search within the netDB the leaseset ls_a through the destination $dest_a$ (we can consider that $dest_a$ is the key and ls_a is the value, in a hash table).

C. I2P file-sharing environment

I2P provides a secure layer for applications to communicate anonymously among themselves. On top of this layer, different file-sharing applications were adapted to work with the concept of *destinations*. Three main file-sharing clients are available within the I2P network: a Gnutella-based named *I2Phex*, an aMule-based called *iMule* and a BitTorrent-like named *I2PSnark*. In this paper, we will focus on the I2PSnark client. We previously showed [1] that this client is the most used client in the I2P file-sharing environment.

III. Group-based characterisation

This section first presents our strategy to perform a group-based characterisation in the I2P network and the monitoring architecture we employ to recover network metadata. Then, it introduces the correlation coefficient used for our analysis.

A. Strategy for characterisation

We consider two variables: the behaviour of I2P users on the system on one side and the behaviour of I2PSnark applications on the other side. Figure 1 shows our objective, considering data from the real network. We take into consideration the number of detected I2PSnark applications and the number of detected users from one city to illustrate our objective. We aim at establishing to what extent this particular set of users contributes to the file-sharing activity detected for this particular period. A positive correlation between these two set of data would allows us to determine that these users are actually performing file-sharing on the network.

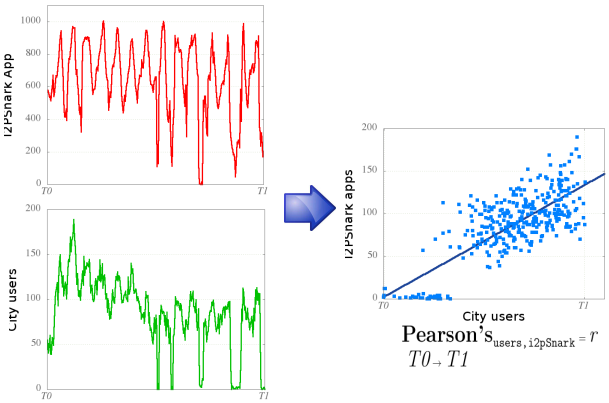

Fig. 1. A correlation strategy for group-based characterisation

Through a *bivariate correlation analysis* we can determine if two variables (I2P users and I2PSnark applications) present a *dependent relationship* and establish whether this dependency is positive or negative. The data we are analysing is of type *ratio* [4]. According to this particular type of data, we consider the *Pearson's correlation coefficient*, which gives us a measure of the *linear dependence* of two variables.

B. Monitoring architecture

In order to determine the number of users from a particular city and the number of I2PSnark applications we employ our distributed monitoring architecture [1]. This distributed architecture consists in a set of monitoring floodfill nodes, distributed over I2P's Kademlia database. These monitoring floodfill nodes are passive, thus undetectable for the rest of the network and collect network metadata, *i.e.* routerinfos and leasesets.

We do not consider a snapshot of the network in a particular moment in time, but rather consider every published routerinfo and leaseset. We aggregate every collected network metadata record in a one-hour period into a single value, thus obtaining the total number of routerinfos and leasesets per hour.

We are considering the geographical localization of I2P users, which is obtained through their routerinfos. Each routerinfo contains, among other parameters, the IP address and the port number where the I2P router[2] can be contacted.

[2]The I2P router is the software that allows Internet users to connect to the I2P network.

We employ a local database based on the MaxMind service determine the country, the region and the city of a given IP address.

C. Pearson's correlation coefficient

Pearson's coefficient provides an output between -1 and 1. According to Cohen [5], a correlation value above 0.50 (or less than -0.50) is considered as strong, a value between 0.30 and 0.50 (or between -0.50 and -0.30) as moderate and finally, a correlation between 0.30 and -0.30 as weak. We only retain positive values in our analysis, since with more I2P users we expect more I2PSnark applications (in the case of a positive correlation).

We need an additional parameter to properly interpret our correlation analysis, the coefficient of determination, given by r^2, where r is Pearson's coefficient. This coefficient is extremely important in our analysis, since it determines to what extent the changes of a set of users are responsible for the changes in the number of detected I2PSnark applications. These changes correspond to the variance of the data and are what we consider the *activity* of users or file-sharing applications. Therefore, the coefficient of determination indicates us to what extent users' activity explains file-sharing activity.

IV. Experimental results

We apply the Pearson's correlation coefficient in order to analyse the relationship among users from a particular city and the number of I2PSnark applications detected. This section first details the setup of our experiment and the methodology of our analysis. Then, we present our monitoring results for the fifteen-day period, where we show the detected number of I2P users and I2PSnark applications. Finally, the section presents different correlation study cases considering three cities and I2P's file-sharing activity.

A. Experiment setup

We monitored the real I2P network from *2013-03-15 CEST* to *2013-03-30 CEST*. We consider that three weekends is a good time window to detect a long-lived correlation between a particular city and I2PSnark applications. We deployed our monitoring architecture over the PlanetLab testbed [6], and due to technical limitations, we used seventy monitor floodfill nodes, which gives us an approximate coverage of 70% of the total network.

We gather 70% of all routerinfos and leasesets within the netDB, where all values stored are *uniformly* distributed over the netDB. Therefore, this partial coverage does not affect our correlation analysis, since the number of routerinfos retrieved is proportional to the number of leasesets retrieved.

B. I2P monitoring results

We measured the number of I2P users and geolocalized them, thus obtaining the hourly number of users from a particular city throughout the fifteen-day period, *i.e.* 360 data points. Figure 2 depicts the number of users from Moscow, Saint Petersburg and Munich detected during our period of analysis.

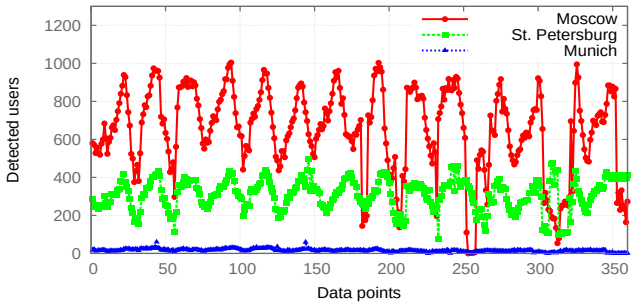

Fig. 2. Number of I2P users from Moscow, Saint Petersburg and Munich

Users from Moscow were the most detected users, with an hourly average of 648 users. Saint Petersburg had an average of 309 users, while Munich has only an average of 16 users per hour. Regarding the number of applications, Figure 3 presents the number of applications detected during the fifteen-day period and the number of I2PSnark clients, which exhibited an hourly average of 87 clients.

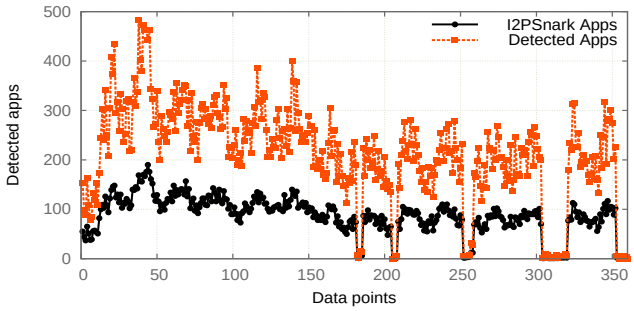

Fig. 3. Number I2PSnark applications detected

C. Methodology

For the following analysis, we first need to determine whether our data fulfils Pearson's data requirements for a correlation analysis, namely data normality, data linearity and data homoscedasticity. Data normality is tested through the analysis of the frequency histogram (users against I2PSnark applications). A *line of best fit* is used to determine whether the distribution of points follows a linear distribution. Homoscedasticity is corroborated by checking the dispersion of points in the plotted data.

Our monitoring architecture covers approximately 70% of the I2P network. This produces few situations where we detect either a low number of users or I2PSnark applications, leading to *outliers*. Outliers [7] are extreme values within our dataset, and can produce data to violate the model's assumptions, such as data normality, producing incorrect results. We only consider extreme-low values in our dataset: we can encounter a low number of I2PSnark applications, but not a high value. *Robust statistical methods* are used to deal with outliers, where we modified a *trimmed estimator* to only analyse the *low* part of the data, *i.e.* those values close to zero. A *trimmed estimator at 10%* only removes every value under the tenth percentile of the ordered data. A trimmed estimator is suitable in our case,

City	Detected users	Overall percentage
Moscow	244223	8%
Saint Petersburg	106688	3.5%
Tokyo	29667	$\sim 1\%$
Yekaterinburg	28507	$\sim 1\%$
Kiev	26262	$\sim 1\%$
Novosibirsk	23090	$\sim 1\%$
Knoxville	17949	$< 1\%$
Paris	14837	$< 1\%$
Berlin	13471	$< 1\%$
Munich	5392	$< 1\%$

TABLE I
MOST ACTIVE CITIES DETECTED

since it discards outliers and keeps only those representative values.

We considered the cities highly active during our measurement, where we detected 16085 cities. Table I depicts these main active cities.

D. Case studies

Moscow and Saint Petersburg presented a high number of users, where the first one contributed to the 8% of the total number of users. We first consider these two cities as case studies. Then, we show that not every active city contributed to the overall file-sharing activity. We illustrate this case with the city of Munich.

1) Moscow: Figure 4 plots our data for Moscow after outliers' exclusion through our modified trimmed estimator.

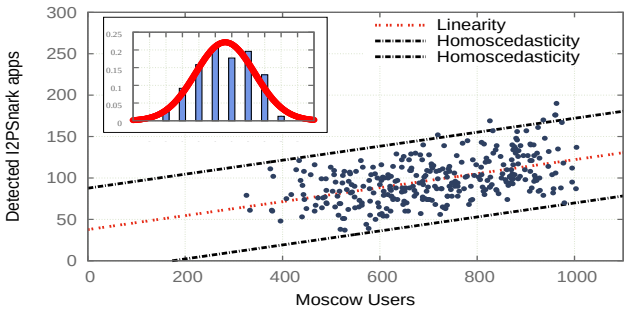

Fig. 4. Data distribution for Moscow/I2PSnark

Data approximately follows a normal distribution and a straight line, presenting the required linearity. Finally, the data keeps a constant dispersion, indicating homoscedasticity. Pearson's coefficient for the fifteen-day sample has a value of $r = 0.4901$, where we can observe a **strong correlation**. The coefficient of determination is $r^2 = 0.2401$, indicating that the activity of users from Moscow explains a quarter of all detected file-sharing activity for this particular fifteen-day period.

2) Saint Petersburg: Figure 5 plots our data for Saint Petersburg after the outliers' exclusion. The results shows that the data complies with Pearson's data requirements. In this case, Pearson's coefficient has a value of $r = 0.3952$, where a **moderate correlation** is observed. The coefficient of determination indicates that the changes in the number of users from Saint Petersburg explain 15.6% of the changes in the number of I2PSnark applications.

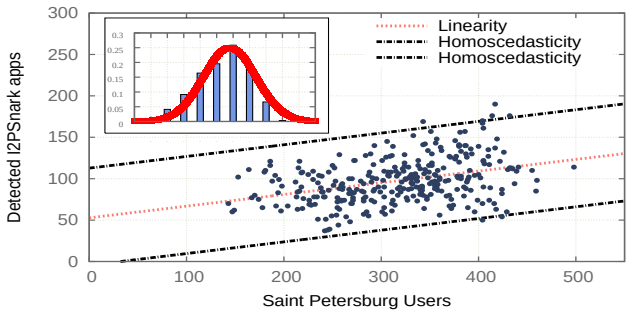

Fig. 5. Data distribution for Saint Petersburg/I2PSnark

3) Munich: We showed that two of the most active cities presented a strong correlation value (for Moscow) and a moderate correlation value (for Saint Petersburg) with I2PSnark applications based on Pearson's coefficient. Let's consider a city like Munich, which had an active daily participation, however it barely contributed to 0.2% of the total number of detected users. Figure 6 plots our data, where a possible correlation is not as clear as with the previous cities. It depicts data after outliers' exclusion where a lack of a normal distribution is observed. Moreover, data presents heteroscedasticity, where for bigger values of users, smaller is the variance. In this case, Pearson's correlation coefficient is not applicable, since data does not comply with the coefficient's data requirements.

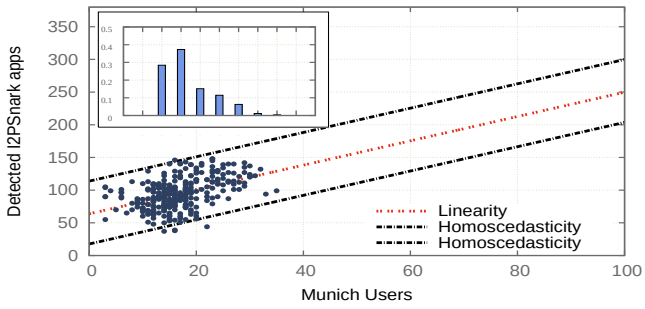

Fig. 6. Data distribution for Munich/I2PSnark

A visual analysis of the data shows that most of the points are concentrated between 14 and 24 users, while the number of detected applications varies from 50 to 150. This behaviour can be observed in the frequency histogram, where both peaks correspond to intervals [14,18) and [18,22) and account for the 65% of the total data points. It indicates that changes in the number of detected I2PSnark applications were not related with changes in the number of users from Munich, leading to conclude that this set of users did not perform significant file-sharing within the I2P network for the studied time period.

V. DISCUSSION

It is essential to consider the privacy and anonymity significance of our results and how they impact on the I2P anonymous network. It is important to distinguish between privacy and anonymity. In simple terms, privacy is the right

to keep personal information from public disclosure, whilst anonymity refers to keep a user's identity hidden.

We did not analyse the content of the anonymous communications and therefore we did not access personal information. Moreover, the type of information in a file-sharing network is not personal, since it is share among all the participants of the network. Thus, we did not jeopardise the privacy of I2P users with our correlation analysis.

It is also important to consider the anonymity provided by the I2P network, how it is affected by our analysis, and the ethics around network monitoring of an anonymous system. Ohm *et al.* [8] showed that there was no special consideration nor *safe harbours* for academic research when conducting any kind of network monitoring. However, and even if there is not any fixed set of rules of thumb, the authors proposed different guidelines to minimise the liability: 1) capture only the data needed for the study; 2) distort the retrieved IP addresses if possible; 3) if sensitive data (IP addresses, for instance) is stored, encrypt it whenever not used; 4) restrict the monitoring to the smallest network required; 5) be aware of filtering tools that might still keep the entire data packet on disk and 6) get a consent from users whenever possible.

We have taken into consideration the guidelines of Ohm *et al.*, with a special focus on obfuscating the IP address retrieved during the monitoring.

VI. Related work

The I2P network is optimised for anonymous hosting and therefore most of the generated traffic remains in the network. Within anonymous systems, the Tor network [2] is the most studied system. However, most of the monitoring techniques applied in Tor [9], [10] can not be applied in the I2P network due to its lack of a central directory or exit nodes. There are two main statistical services for the I2P network. The first service[3] provides approximate values for the number of users in the network and the number of applications deployed. However, this service does not provide the type of applications in the network nor the geographical distribution of the users and therefore there is no characterisation of the users. The second service provides only uptime statistics for I2P's anonymous web sites[4] and does not present any characterisation of the users deploying these web sites.

To our knowledge, there are no analyses within the I2P network towards *group-based characterisation* where users' behaviours is considered.

VII. Conclusion

We have presented the first approach to successfully perform a group-based characterisation in the I2P anonymous file-sharing environment. We showed that despite a strong underlying anonymizing layer, it is possible to analyse users' activities and determine whether their behaviour presents similar patterns with anonymous applications. By accordingly applying Pearson's correlation coefficient, we are able to *link* which cities are the greater contributors to the overall I2P's file-sharing activity during a particular period.

Our previous results [1] showed that there are more than 16000 active cities in the I2P network. In that work, we demonstrated that the activity of two cities representing the 11.5% of all I2P's user-base, explains 40% of all I2P's file-sharing activity. This clearly show that despite the worldwide distribution of I2P users, the two Russian most important cities remain responsible for a considerable share of all anonymous file-sharing activity.

Our perspectives have two axes. In the first place, we aim at automatizing our approach to perform a correlation analysis with a wider set of active cities. In the second place, we need to consider longer periods of analysis, which would allow us to link a smaller set of users to I2P's file-sharing activity. These longer periods will enable us to determine the *trends* within I2P's file-sharing environment, such as which set of users are the one consuming new content first or if the same cities are always the most content consuming cities.

[3]http://stats.i2p.in. Last visited on 10/2013.

[4]tino.i2p.in. Last visited on 10/2013.

References

[1] Juan Pablo Timpanaro, Isabelle Chrisment, and Olivier Festor. A Bird's Eye View on the I2P Anonymous File-sharing Environment. In *NSS 2012*, China, 2012.

[2] Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor: the Second-Generation Onion Router. In *USENIX Security 2013*, San Diego, CA, 2004.

[3] Petar Maymounkov and David Mazières. Kademlia: A Peer-to-Peer Information System Based on the XOR Metric. In *IPTPS 2002*, Cambridge, MA, USA, 2002.

[4] Stanley Stevens. On the Theory of Scales of Measurement. *Science*, 103(2684):677–680, 1946.

[5] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge Academic, 2 edition, January 1988.

[6] Larry Peterson, Tom Anderson, David Culler, and Timothy Roscoe. A Blueprint for Introducing Disruptive Technology into the Internet. In *Proceedings of HotNets–I*, Princeton, New Jersey, October 2002.

[7] Frank Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17-21, 1973.

[8] Douglas C. Sicker, Paul Ohm, and Dirk Grunwald. Legal issues surrounding monitoring during network research. In *Proc. of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, pages 141–148, New York, NY, USA, 2007. ACM.

[9] Damon Mccoy, Kevin Bauer, Dirk Grunwald, Tadayoshi Kohno, and Douglas Sicker. Shining Light in Dark Places: Understanding the Tor Network. In *PETS 2008*, Leuven, Belgium, 2008.

[10] Karsten Loesing, Steven J. Murdoch, and Roger Dingledine. A case study on measuring statistical data in the Tor anonymity network. In *Proc. of the 14th international Conference on Financial Cryptograpy and Data Security*, FC '10, pages 203–215, Berlin, Heidelberg, 2010. Springer-Verlag.