Discovering Patterns of Interest in IP Traffic
Using Cliques in Bipartite Link Streams

Tiphaine Viard\textsuperscript{1}, Raphaël Fournier-S’niehotta\textsuperscript{1}, Clémence Magnien\textsuperscript{2}, Matthieu Latapy\textsuperscript{2}

\textsuperscript{1}CEDRIC CNAM, F-75003 Paris, France
\texttt{tiphaine.viard@cnam.fr, fournier@cnam.fr}

\textsuperscript{2}Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7606, LIP6, F-75005, Paris, France
\texttt{clemence.magnien@lip6.fr, matthieu.latapy@lip6.fr}

Abstract. Studying IP traffic is crucial for many applications. We focus here on the detection of (structurally and temporally) dense sequences of interactions, that may indicate botnets or coordinated network scans. More precisely, we model a MAWI capture of IP traffic as a link streams, \textit{i.e.} a sequence of interactions \((t_1, t_2, u, v)\) meaning that devices \(u\) and \(v\) exchanged packets from time \(t_1\) to time \(t_2\). This traffic is captured on a single router and so has a bipartite structure: links occur only between nodes in two disjoint sets. We design a method for finding interesting bipartite cliques in such link streams, \textit{i.e.} two sets of nodes and a time interval such that all nodes in the first set are linked to all nodes in the second set throughout the time interval. We then explore the bipartite cliques present in the considered trace. Comparison with the MAWILab classification of anomalous IP addresses shows that the found cliques succeed in detecting anomalous network activity.

1 Introduction

Attacks against online services, networks, and information systems, as well as identity thefts, have annual costs in billions of euros. Network traffic analysis and anomaly detection systems are of crucial help in fighting such attacks. In particular, much work is devoted to the detection of anomalous patterns in traffic. This work mostly relies on pattern search in (sequences) of graphs to, which poorly captures the dynamics of traffic, or on signal analysis, which poorly captures its structure. Instead, we use here the \textit{link streams} framework, which allows us to model the both structural and temporal nature of traffic in a consistent way. Within this framework, a clique is defined by a set of nodes and a time interval such that all the nodes in that set continuously interact with each other during this time interval. Such patterns may be the signature of various events of interest like botnets, DDoS, and others.

We present in Section 2 our link stream modeling of traffic captures, that calls for specific notions as this traffic is bipartite. We present in Section 3 fast heuristics for finding cliques of interest in practice. We briefly present in Section 4 the MAWI dataset that we use for our experiments, and we detail in Section 5 the results of these experiments. Related work is overviewed in Section 6 and we discuss some perspectives in Section 7.
2 Traffic as a Link Stream

IP traffic is essentially composed of packets, each with a source address and a destination address. Routers forward these packets towards their destination, and one may capture traffic traces by recording the headers of packets managed by a router, together with the time at which they manage them. Such traces are generally collected on firewalls, access points, or border routers. As a consequence, they often have a bipartite nature: they capture exchanges between two disjoint sets of devices (for instance the ones in a company LAN and the outside internet), as these routers are not in charge of traffic between devices within one of these sets. This leads to the following definition of bipartite link streams, that extends the classical definitions of bipartite graphs [1] and of link streams [2].

A bipartite link stream $L = (T, \top, \bot, E)$ is defined by a time span $T$, a set of top nodes $\top$, a set of bottom nodes $\bot$, and a set of links $E \subseteq T \times \top \times \bot$. If $(t, u, v) \in E$ then we say that $u$ and $v$ are linked at time $t$. We say that $l = (b, e, u, v)$ is a link of $L$ if $[b, e]$ is a maximal interval of $T$ such that $u$ and $v$ are linked at all $t \in [b, e]$. We call $l = e - b$ the duration of $l$. See Figure 1 for an illustration. We consider undirected links only: we make no distinction between $(t, u, v)$ and $(t, v, u)$ in $E$.

In order to model traffic as a link stream, we need to transform packets exchanged at precise timestamps into continuous interactions. To do so, we consider that two IP addresses are continuously linked together from time $b$ to time $e$ if they exchange at least one packet every second within this time interval. This leads to the following definition of $E$: it is the set of all $(t, u, v)$ such that $u$ and $v$ exchanged a packet at a time $t'$ such that $|t' - t|$ is lower than half a second. See Figure 2 for an illustration.
Fig. 2: Left: packet exchanges at precise timestamps. Right: the link stream obtained by considering that two nodes are continuously linked together if they exchange at least one packet every second.

3 Finding Cliques of Interest

In a bipartite graph $G = (\top, \bot, E)$, a clique is a couple $(C_\top, C_\bot)$ with $C_\top \subseteq \top$ and $C_\bot \subseteq \bot$ such that $C_\top \times C_\bot \subseteq E$. In other words, there is a link between each element of $C_\top$ and each element of $C_\bot$.

We define similarly a clique in a bipartite link stream as a tuple $(C_\top, C_\bot, I)$ with $C_\top \subseteq \top$, $C_\bot \subseteq \bot$ and $I$ an interval of $T$ such that $I \times C_\top \times C_\bot \subseteq E$. In other words, each element of $C_\top$ is linked to each element of $C_\bot$ for the whole duration of $I$. We call $|C_\top \cup C_\bot|$ the size of the clique and $|I|$ its duration. A clique is maximal if there is no other clique $(C'_\top, C'_\bot, I')$ such that $C_\top \subseteq C'_\top$, $C_\bot \subseteq C'_\bot$ and $I \subseteq I'$. See Figure 3 for an illustration.

Fig. 3: Two maximal cliques of the bipartite link stream of Figure 1: $(\{u, v\}, \{a, b\}, [3, 5])$ (in blue) and $(\{u, v\}, \{a\}, [8, 10])$ (in green).

As explained above, we explore cliques as patterns of interest in IP traffic. It is easy to extend to the bipartite case the (non-bipartite) link stream clique computation algorithm presented in [4] (see below). However, we face two issues. First, clique computations are costly, and enumerating all cliques is out of reach in our case; we therefore resort to sampling. Second, not all maximal cliques are interesting: in particular, a node and all its neighbors (which may be very numerous) always form a clique in the bipartite case, which has little interest for us.
As a consequence, we focus here on balanced cliques: a clique \((C_T, C_I, I)\) is balanced if and only if the difference in size between \(C_T\) and \(C_I\) is at most 1. We then search for balanced cliques as follows. We iteratively build \((C_T, C_I, I)\) from the empty clique of maximal duration, \((\emptyset, \emptyset, T)\). At each step we choose a random node \(v\) in \(\top\) or \(\bot\) such that \(v\) is linked to all nodes in \(C_I\) or \(C_T\) respectively, during an interval \(I_v \subseteq I\), and we update the current clique into \((C_T \cup \{v\}, C_I, I_v)\) or \((C_T, C_I \cup \{v\}, I_v)\), respectively. In order to ensure the obtained clique is balanced, we alternatively choose \(v\) in \(\top\) and in \(\bot\) at each iteration. We stop when no node can be added.

This process samples a maximal balanced clique, and we run it many times to obtain a large set of such cliques (see Section 5). It must be clear that this set may be biased by the sampling process, which is not uniform. However, since our primary goal is to explore the relevance of some cliques in the context of IP traffic, studying this bias is out of our scope. Notice also that the found clique is maximal amongst the set of balanced cliques, but not necessarily maximal in the stream: it may indeed be included in an unbalanced clique.

![Diagram of balanced clique](image)

Fig. 4: From the blue clique \((\{u, v\}, \{a\}, [1, 10])\) our algorithm builds the clique \((\{u, v\}, \{a, b\}, [4, 5])\) by adding \(b\), which reduces the time span from [1, 10] to [4, 5].

Our sampling process tries to add as many nodes to the clique as possible, which generally induces a reduction of its duration, see Figure 4. As we are both interested in large and long balanced cliques, we include in the sampled set all the intermediary maximal balanced cliques built during the process.

4 Dataset

We use here a one hour IP traffic capture from the MAWI archive, more precisely from the A Day in the Life of the Internet (DITL)\(^1\) initiative, from 2013, June, 24\(^{\text{th}}\), 23:45 to 2013, June 25\(^{\text{th}}\), 00:45. This trace lasts 3,600 seconds, during which 88,266,534 packets are sent involving 992,466 nodes. The topology is represented in Figure 5: 408,751 nodes are part of WIDE, and 583,715 are outside of WIDE. WIDE/Non-WIDE sets are nearly balanced, which improves our chances of finding large balanced cliques.

\(^1\) [http://mawi.wide.ad.jp/mawi/ditl/ditl2013](http://mawi.wide.ad.jp/mawi/ditl/ditl2013)
We transform the data into a bipartite link stream \( L = (T, \top, \bot, E) \), where \( T = [0, 3600] \), \( \top \) contains all observed WIDE IP addresses, and \( \bot \) contains all other observed IP addresses. As explained in Section 2, \( E \) is the set of links obtained by considering that nodes interact for one second every time they exchange a packet. In contains 6, 206, 295 links.

In addition to this raw data, we use the MAWILab database \([5]\), which gives labels locating traffic anomalies in the MAWI archive. These labels are obtained using an advanced graph-based methodology that compares and combines different and independent anomaly detectors. This database indicates that there is a total of 488 anomalous IP addresses in our dataset, that we will use to interpret the results in the following.

5 Results

We ran our sampling algorithm on the MAWI dataset. We ran 14 instances in parallel on a server\(^2\) for 106 days (3 months and 17 days). This led to a total of 1,291,084,661 sampled cliques, among which there were a great number of duplicates: 198, 718, 323 are distinct maximal balanced cliques, which we study in the rest of the paper.

We call anomalous clique a clique which contains at least one IP address that is flagged as anomalous in the MAWILab database.

We show in Figure 6 the distribution of sampled clique sizes and the inverse cumulative distribution of their durations. Many are of size 2 or 3, which has little interest: cliques of size 2 are single links, and bipartite cliques of size 3 are just composed of a node linked to two other nodes. Since any node with \( k \) neighbours leads to \( k(k - 1)/2 \) balanced 3-cliques, the large number of such cliques is unsurprising and brings no significant information. We therefore focus on cliques of size 4 or more. Our sample contains 275, 647 such cliques.

\(^2\) A linux machine with 24 cores at 2.9 GHz and 256 GB of RAM.
The sampled cliques of 4 nodes or more involve 29,744 distinct nodes, 94 of which anomalous. The fraction of anomalous nodes in these cliques therefore is $3.1 \cdot 10^{-3}$, much larger than in the whole dataset, $5 \cdot 10^{-4}$. This indicates that maximal balanced cliques are related to anomalous activity, as suspected.

While most maximal balanced cliques have a duration close to 1 second, the duration distributions show that there are very long cliques. However, duration is highly influenced by size, as explained above, and so we display the duration distribution for each clique size separately in Figure 7 (for readability, we show them in two plots). As expected, the duration of large cliques is in general shorter than for cliques involving only few nodes. The 18 cliques spanning more than 500 seconds all have 4 or 5 nodes, the longest being of size 4 and duration 1045.75 seconds. Still, there are cliques involv-
Fig. 7: Inverse cumulative distribution of durations of sampled balanced cliques of a given size. Top: for sizes from 4 to 7 nodes. Bottom: for sizes from 8 to 12 nodes.

Importantly, among the 18 cliques spanning more than 500 seconds, 17 are anomalous. Going further, Figure 8 shows that more than 80% of all cliques involving at least 4 nodes during at least 8 seconds are anomalous. This demonstrates the fact that long non-trivial cliques may be used to detect anomalous activity.

We deepen our understanding of large cliques by displaying in Figure 9 the time span of all maximal balanced cliques of size 4 or more: we sort cliques according to their starting time, which gives their vertical position in the plot. We then represent each clique \((C_1, C_2, I)\) by an horizontal line from the beginning of \(I\) to its end. Colors indicate clique size. This plot confirms our previous observations; it shows the preva-
Fig. 8: Fraction of anomalous cliques (vertical axis) having duration greater than a given value (horizontal axis). Bottom: for all cliques. Top: for cliques involving at least 4 nodes.

lence of smaller cliques among the longest ones, and it shows that we succeed in finding significant cliques during the whole time span of the dataset.

Fig. 9: Representation of the temporal features of our balanced cliques of size 4 or more. Each clique \( (C_T, C_L, I) \) is represented by a horizontal line of length \( |I| \) with a color representing its size (color code on the right). These lines are ranked according to their beginning time.
However, it also displays a sharp increase shortly after time 3000, corresponding to a large number (31, 201) of short balanced cliques that start then. Figure 10 displays these cliques in more details, and shows that they almost all involve anomalous nodes. This confirms that clique structures are related to anomalous activity.

Interestingly, it seems that these anomalies cannot easily be distinguished from other anomalies directly on simple plots like the number of distinct nodes or links over time, see Figure 11. Although there is a peak in both plots at time 3060, it is not different from other peaks, yet it is the only one corresponding to such a sharp increase in Figure 9. This indicates that cliques highlight specific features of this event, that are worth investigating further.

We therefore display in Figure 12 the graph induced by the balanced cliques we found at seconds 3059, 3060, or 3061. The nodes of this graph are the nodes involved in at least one of these cliques, and they are linked together if this link occurs in at least one of these cliques. The graph has a large connected component with two anomalous nodes linked to 228 distinct non-anomalous nodes, confirming that most of these cliques are the signature of a same event (they actually are parts of a much larger but unbalanced clique). This signature is typical of a coordinated scan in a class C network.

6 Related Work

IP traffic has been extensively studied for decades with very powerful approaches relying on signal processing and/or machine learning [6,7,8]. Graph-based approaches are also appealing, and were used previously. In [9] for instance, the authors use graphs
to represent temporal dependencies in traffic and characterize traffic behavior. In [10], the authors model the traffic as a bipartite graph, then used one-mode projections and clustering algorithms to discover end-host and application behavior clusters. Clustering communication behavior was also proposed in [11], where authors discusses the relevant features before analysing significant nodes for long periods.

Graph-based approaches are however limited in their ability to capture temporal information, which is crucial for traffic analysis. Indeed, they generally rely on splitting data into time slices, and then aggregate traffic occurring into each slice into a (possibly weighted, directed, and/or bipartite) graph. One obtains this way a sequence of graphs, and one may study the evolution of the properties of these graphs, see for instance [12]. However, choosing small time slices leads to graphs that are almost empty and bring little information, if any. Conversely, large slices lead to important loss of
information as the dynamics within each slice is ignored. As a consequence, choosing appropriate sizes for time slices is extremely difficult and it is a research topic in itself [13,14]. There is currently an important interdisciplinary effort for solving these issues by defining formalisms able to deal with both the structure and dynamics of such data. The link stream approach is one of them [5], as well as temporal networks and time-varying graphs [15,16,17]. Up to our knowledge, these other approaches have not yet been applied to network traffic analysis.

Our work is also related to the search of motifs in a temporal network, see for instance [18,19]. These works however focus on enumerating 2-edges or 3-edges motifs, which is a complementary to our work, and would certainly provide interesting insight on traffic too.

7 Discussion

We have shown that cliques in bipartite link streams modeling of IP traffic allow to detect anomalous activity in IP traffic: long cliques of significant size involve anomalous nodes known from MAWILab, although they are rather small, and apparition of many small cliques at a same time indicate coordinated activity like distributed scans. This work however only is a first step, and it raises many questions.

In particular, the computational cost of our method is prohibitive. Algorithmic work is therefore needed to design faster clique detection heuristics, and to search for quasi-cliques. One may also preprocess the stream by iteratively removing nodes of degree 1, which represent a large fraction of the whole and cannot be involved in non-trivial cliques. Going further, one may use the link stream modeling to define many other structures of interest regarding anomalous traffic. Exploring other modeling assumptions is also appealing, in particular the fact that we linked nodes together if they exchanged
packets at least every second (other time limits may be interesting), the fact that we considered undirected links whereas traffic is directed, or the use of port or protocol information present in the data.

References

1. Latapy, M., Magnien, C., Vecchio, N.D.: Basic notions for the analysis of large two-mode networks. Social Networks 30(1) (2008) 31–48
2. Latapy, M., Viard, T., Magnien, C.: Stream graphs and link streams for the modeling of interactions over time. https://arxiv.org/abs/arXiv:1710.04073 (2017)
3. Viard, T., Latapy, M., Magnien, C.: Computing maximal cliques in link streams. Theor. Comput. Sci. 609 (2016) 245–252
4. Himmel, A., Molter, H., Niedermeier, R., Sorge, M.: Enumerating maximal cliques in temporal graphs. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM. (2016)
5. Fontugne, R., Borgnat, P., Abry, P., Fukuda, K.: Mawilab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking. In: ACM CoNext '10. (December 2010)
6. Borgnat, P., Dewaele, G., Fukuda, K., Abry, P., Cho, K.: Seven years and one day: Sketching the evolution of internet traffic. In: INFOCOM 2009. 28th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 19-25 April 2009, Rio de Janeiro, Brazil, IEEE (2009) 711–719
7. Fontugne, R., Abry, P., Fukuda, K., Borgnat, P., Mazel, J., Wendt, H., Veitch, D.: Random projection and multiscale wavelet leader based anomaly detection and address identification in internet traffic. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015, IEEE (2015) 5530–5534
8. Himura, Y., Fukuda, K., Cho, K., Borgnat, P., Abry, P., Esaki, H.: Synoptic graphlet: Bridging the gap between supervised and unsupervised profiling of host-level network traffic. IEEE/ACM Trans. Netw. 21(4) (2013) 1284–1297
9. Asai, H., Fukuda, K., Abry, P., Borgnat, P., Esaki, H.: Network application profiling with traffic causality graphs. Int. Journal of Network Management 24(4) (2014) 289–303
10. Xu, K., Wang, F., Gu, L.: Behavior analysis of internet traffic via bipartite graphs and one-mode projections. IEEE/ACM Trans. Netw. 22(3) (2014) 931–942
11. Jakalan, A., Jian, G., Zhang, W., Qi, S.: Clustering and profiling IP hosts based on traffic behavior. JNW 10(2) (2015) 99–107
12. Latapy, M., Hamzaoui, A., Magnien, C.: Detecting events in the dynamics of ego-centred measurements of the internet topology. J. Complex Networks 2(1) (2014) 38–59
13. Benamara, L., Magnien, C.: Estimating properties in dynamic systems: the case of churn in P2P networks. In: International Workshop on Network Science for Communication Networks (NetSciCom 2010), In Conjunction with IEEE Infocom. (2010)
14. Leo, Y., Crespelle, C., Fleury, E.: Non-altering time scales for aggregation of dynamic networks into series of graphs. In: Proc:ACM Conference on Emerging Networking Experiments and Technologies, CoNEXT. (2015)
15. Casteigts, A., Flochini, P., Quattrociocchi, W., Santoro, N.: Time-varying graphs and dynamic networks. JPEDS 27(5) (2012) 387–408
16. Wehmuth, K., Ziviani, A., Fleury, E.: A unifying model for representing time-varying graphs. In: 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, October 19-21, 2015, IEEE (2015) 1–10
17. Holme, P.: Modern temporal network theory: a colloquium. The European Physical Journal B 88(9) (2015) 1–30
18. Paranjape, A., Benson, A.R., Leskovec, J.: Motifs in temporal networks. In: Proc. ACM International Conference on Web Search and Data Mining, WSDM. (2017)
19. Kovanen, L., Karsai, M., Kaski, K., Kertész, J., Saramäki, J.: Temporal motifs in time-dependent networks. Journal of Statistical Mechanics: Theory and Experiment 2011(11) (2011) P11005