Cyberspace and Time: A Light through the Fog?

AUTHOR
Rémi GÉRAUD

ABSTRACT

One usually describes cyberspace as a virtual realm. In truth, it is grounded in real infrastructures, which therefore inherit local regulations and broader geopolitical influences from the location where they are set up. The consequences of this attachment to a place can be measured to some extent from the communications happening in cyberspace, i.e., through communication networks. But there is also a temporal anchorage resulting from phenomena such as synchronisation in communications which, when analysed, reveals deep and surprising proximities, shedding new light on the relationships between actors in cyberspace. We discuss in particular how an analysis of synchronicity in cyberspace helped to identify collusions and a diffusion network active in spreading radical messages, as well as providing insight in these groups' internal structure.

KEYWORDS

Territorial sciences, Cyberspace, Internet, Synchronicity, Mathematics

RÉSUMÉ

Le cyberespace constitue un domaine généralement décrit comme virtuel. Mais il est bâti en réalité sur des infrastructures tout à fait existantes, qui sont dès lors sujettes aux législations territoriales, et aux influences géopolitiques héritées des lieux où elles sont installées. Cet ancrage terrestre laisse des traces, mesurables, dans les échanges qui se tiennent dans le cyberespace, *via* les réseaux de télécommunication. Mais il y a également un ancrage dans le temps, réalisé par exemple par la synchronicité des échanges, et qui révèle des proximités d'une autre nature, éclairant les relations entre acteurs du cyberespace selon un angle nouveau. En particulier, nous illustrons comment une étude du temps dans le cyberespace a permis d'identifier des collusions à large échelle dans la diffusion de messages à contenu radical, et la structure des groupes de diffusion.

MOTS CLÉS

Sciences du territoire, cyberespace, internet, synchronicité, mathématiques

INTRODUCTION

The fabric of cyberspace is woven from telecommunication channels. Over these channels, information is created, transported, transformed, and exchanged; but these channels are drawn between concrete devices and infrastructures: this makes them objects of real geopolitical interest. Indeed these infrastructures are bound to undergo various cultural, political, and socio-economic pressures, which therefore bear in some way on any network in which they participate (Douzet *et al.*, 2014).

One may thus wonder to what extent it is possible to recognise or identify such influences, mapping similarities or dissimilarities, by participating in a communication network. Indeed, the network itself introduces additional information along with the mere messages it transports; for instance: routing and addressing protocols (BGP, ARP...) used by the different devices in the network to organise themselves, regulate traffic, avoid congestion, and other such tasks;

incomplete or partial messages, echoes, resulting from, e.g., some defective piece of equipment; and side channel information, that can be measured from the network but is not explicitly transported over it, such as latency, packet drop, or timing correlations.

In this work, we discuss how to capture and leverage timing information on a public network to define a first notion of synchronicity, and form rough communities based on activity patterns; we then refine this analysis to identify message diffusion patterns (causal avalanches). Finally, we use network latency estimates to identify automated posters on the public network, and show that they are likely to be operated by few entities.

All in all, we provide a framework and tools to detect non-local collusions, which may be attributed to forethought, automation, or hidden communication channels not immediately accessible to the analyst, but visible through their effects –all three possibilities being relevant to a thorough understanding of cyberspace.

1. WHY TIME MATTERS

1.1. The network fog of war

Computer networks are built atop infrastructures that belong to the physical world, where cyberspace collides with geographic space. Because of this, key assets (datacenters, largest network routes, etc.) can be mapped, to some extent, using traditional means, and lend themselves to an extended geopolitical analysis (Douzet *et al.*, 2014; Limonier, 2014). The resulting cartography is still in its infancy, although it shows beyond doubt that some states have already embraced cyberwarfare as part of their doctrine.

However, building (and concealing) new infrastructures is a long and costly process. While this is compatible with some objectives, e.g. controlling data routes or storage, it constitutes too much of a risk of being exposed in sensitive operations, such as influence campaigns or massive distributed denial of service attacks. This does not mean that state actors refrain from engaging in such operations; but they need plausible deniability.

To that end, they can leverage the absence of integrity and authenticity guarantees offered by computer networks, in particular Internet[1]. Indeed, attackers can send crafted packets having wrong sender information (which is one way to launch DDoS attacks), or hide behind intermediation platforms such as social media websites. As a result, attack attribution is made very difficult, and identifying the attack's precise source is close to impossible. This is a famous issue in the context of network security, which ridicules attempts at automatically retaliating in name of "active defence" strategies, lest some additional information is known from e.g. intelligence sources that the attack indeed originated from the claimed source. Thus, attackers can craft data to hide their position, number, and possibly any sensitive information, thereby providing strong deniability guarantees.

What we suggest in this note is that the *immediacy* of computer network communications, made sharper by technological evolution, provides the analyst with new means to measure "closeness" or "similarity" beyond geographical vicinity, and complementary to it.

[1] By "Internet" we really mean the Internet protocol (IP), and its two most used super-layers: the transport control protocol (TCP/IP) and the user datagram protocol (UDP/IP). Together they account for the entirety of the traffic to web pages and to more than 80% of all traffic worldwide.

From a theoretical standpoint, we hope to extend the toolkit available for researchers and actors to analyse abstract networks (and in particular communication networks), aiming at extending the classical cartographic approach to the virtual world.

1.2. Synchronicity, causality, sychnometry

This "geography of time" is not a mere accident of the network's internals, but reveals underlying structures, and to a degree, causal influence between actors. As such it provides a better understanding of event dynamics in cyberspace, and gives additional insight into the geopolitical approaches of cyber actors.

We define this notion of time vicinity using the mathematical notion of stochastic processes, which give a precise meaning to "random" events occurring over time, and give means to measure their statistical dependence. There could be many ways in which two events influence one another; we focus on two of them: excitation and repetition. Excitation means that a certain class of events suddenly becomes much more probable; informally, this captures synchronicity between events (including runaway phenomena). Repetition means that the occurrence of an event strongly correlates to the occurrence of the same event at some later time; we refer to this analysis as sychnometry[2].

It is a basic fact that statistical correlation usually does not imply causation, and thus we shall refrain from making such a confusion. However, under some hypotheses, we can make an educated guess about the causal structure of events. For instance, if event A happens before event B for all observers, then B is unlikely to be the cause of A. Alternatively, if we observe that causation typically happens over a short period of time (e.g., days), but A and B are separated by a too short (e.g., millisecond) or too long (e.g., months) amount of time, then it is unlikely that they are causally (directly) related. Mathematically, we model the phenomena of interest as Hawkes processes (Bacry *et al.*, 2015), fitted using Maximum Likelihood estimation; from this we can extract synchronicity and sychnometric information efficiently, and to some extent reconstruct the (early) causal structure of events. We shall insist, here and later, that this is only causality in a probabilistic sense, and therefore may return wrong results (as any other tool).

The results of this analysis on a concrete case are given in section 3 below, after some technical preliminaries and methodological care. The context of heated debate over cyberspace territoriality and sovereignty is paramount to understand our motivations; we expect that the approach highlighted here participates in mending the disconnect between "virtual" and "concrete" geopolitics.

Our focus is on communication networks in general, of which social networks are an example and provide an interesting source, as they happen to be central hubs of influence for geopolitical questions.

2. MEASURING TIME OVER THE INTERNET

2.1. Defining and measuring network latency

Over the Internet, data is grouped into "packets" before transmission, that are transported from network device to network device until their destination is reached, or the packet is lost or

[2] From Greek "συχνός", which means "frequent, repeated".

corrupted to the point that transport is no longer relevant. In theory, packets are forwarded by intermediate network nodes asynchronously (first-in, first-out), which results in a processing delay, a transmission delay (pushing the packet's bits onto the link), and a propagation delay, i.e., the time taken for a signal to reach its destination.

In reality, additional mechanisms are used to shape traffic, guarantee delivery or fairness between users (e.g., "quality of service" mechanisms), and the physical medium's limitations must be taken into account (e.g., in wireless networks). All of these result in varying queuing delays that add to the full transmission time (Nygren *et al.*, 2010).

The round-trip delay time (RTT) accounts for all the above effects and measures the time taken between the moment a message is sent and the moment it is fully echoed back. However, the RTT does not discriminate between forward (from sender to recipient) and backward (from recipient back to sender) delays. A more robust approach is to use active measurements, for instance with the minimum-pairs algorithm (Abdou *et al.*, 2015), which requires the use of three trusted and synchronised nodes to reliably measure the one-way latency to a fourth (not necessarily trusted) node. Alternative methods are possible but typically less reliable (Gummadi *et al.*, 2002).

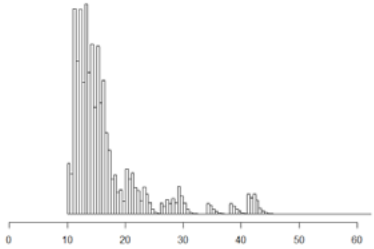
Note that any measurement performed on a remote device will necessarily be stained by some latency (this is akin to the situation in astrophysics, as one observes a remote star). This has to be corrected, and can be to some degree, as we now explain.

2.2. Time-of-flight measurements and relative positioning

Informally, the more intermediate nodes a packet has to go through, the longer the delay. By probing several well-positioned nodes on the network we can therefore infer how "far away" they are in terms of network hops, and to some extent reconstruct the underlying "informational highways", even if they are invisible from available (meta)data.

Representing this information is highly non-trivial however, as latency measurements aggregate several independent delay sources, and vary over repeated experiments. Nevertheless, large trends are visible and this average "time-of-flight" already highlights privileged or hampered routes. In particular we are interested in the latency distribution, which contains information about typical delays and variations (fig. 1)[3].

Figure 1. Latency distribution to a Twitter CDN node. Horizontal scale is in milliseconds



[3] Time measurements are precise to the millisecond; variations are not due to any "measurement error" but to the superposition of various networks mechanisms delaying packet transmission.

2.3. Trusted time reference

In some situations we can have a public and trusted time reference. To that end we synchronise with GPS time (Lombardi *et al.*, 2001) and measure latency distribution from a fixed point to a given public server. In our case, the public server will be one of Twitter's CDN nodes. This allows us to measure the time elapsed between the moment a message is sent to the Twitter platform and the value indicated by the timestamp appearing on the platform when that message has been received. By monitoring the public API, we therefore get precise timings of when we received notification of the message, when the notification was sent (corrected by subtracting the one-way delay estimation from the server to our device), and when the message was acknowledged on the Twitter platform (as per the API-provided timestamp). Comparing these timings shows that we get a reasonably precise estimate of when the message is received by Twitter, and we will henceforth proceed using this information.

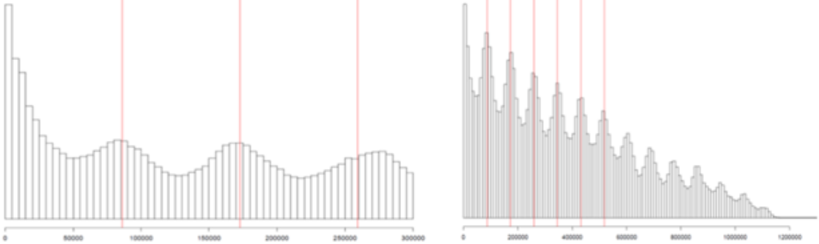
Using a trusted time reference, we can monitor events, record information, and order events in an unambiguous way. This was done for Twitter data over the course of several months. We can then use the tools described in section 1, which gives the results described below.

3. SYNCHRONICITY PATTERNS AND NON-LOCAL COLLUSIONS

3.1. Synchronicity measures: Multi-scale visualisation of influence

On figure 2, we represent the synchronicity/excitation measures extracted from retweet activity based on our captured data. A striking observation is that there is a regular pattern of activity strongly aligned with a given time zone (despite users declaring various time zones), which is observed independently from the message content. These graphs show, for instance, that a message is more likely to be retweeted 24 hours after the initial post, than 12 hours after. The near linear (rather than exponential) decay in synchronicity may be interpreted as the manifestation of a *causal avalanche*: a first set of users propagate the message over the next day, then another set of users (following roughly the same activity pattern) propagate the message over the next day, etc. Such a scenario can be simulated and exhibits a similar linear decay in synchronicity. However this information alone is not sufficient to guarantee that what is observed actually follows a propagation avalanche.

Figure 2. Synchronicity measures for retweets over the course of 3 days (left) and two weeks (right)

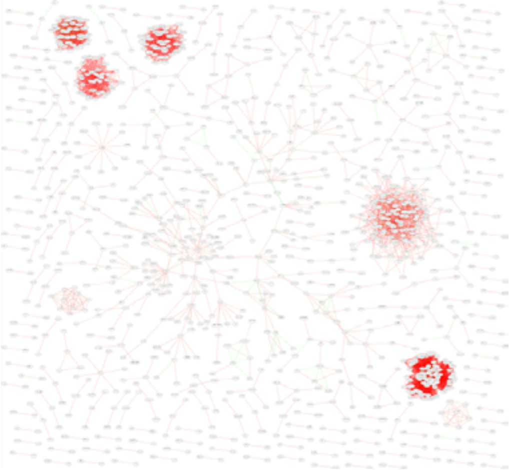


Vertical red lines indicate 24-hour periods. This graph shows strong correlations with the day-night cycle corresponding to a geographic time zone, a near-linear decrease in synchronicity over time.
Horizontal scale is in seconds from a reference point, as explained in section 2

Identifying the peaks in figure 2 is easily done algorithmically as it corresponds to values where the first derivative vanishes and the second derivative is negative. We should insist here that figure 2 is obtained without any smoothing and is the direct result of synchronicity measures over aggregated data, which shows how very clear and reliable activity patterns can be identified.

Limiting ourselves to the first minutes of diffusion, synchronicity measures provide a representation of how information propagates across the network, from neighbour to neighbour (fig. 3).

Figure 3. Synchronicity measures, capped to 3 minutes,
mapped as relationships between nodes on a portion of the Twitter network



Two nodes in this graph are linked when they typically react within 3 minutes of one another

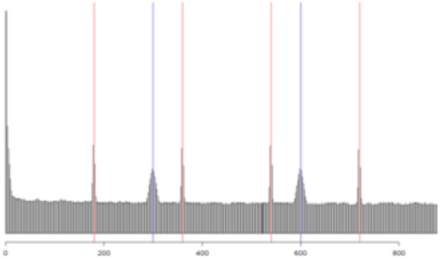
Synchronicity can be observed at several scales, and we have very good temporal resolution (1 ms). This gives a multi-scale picture of influence in the network, complemented by other community measures (e.g., network centrality).

3.2. Sychnometry measures: Visualising collusions

Our other measure focuses on repetition in activity, even when there is no link between two users (i.e., they do not interact). Following probability theory, we should expect a uniform ("flat") distribution for synchrometric estimates when all the nodes behave independently. As we saw on the synchronicity measures, there is latent variable in the form of a day-night cycle that must be accounted for. But in fact, when we turn our attention to very short time spans (where the day-night cycle's influence is negligible), and a portion of the network known to contain radical messages, we obtain figure 4.

What is visible on this figure is the presence of *highly coordinated actors*: actors that post at different moments of the day, and do not interact with one another, but do so "similarly'. They would not be otherwise visible! This is remarkable not only because of this sudden break through the mist, but because the sychnometric patterns appear extremely thin. Concretely, this means that groups of posters follow the exact same pattern within less than a second. They can be automatically identified using classical outlier detection techniques. Our interpretation is that this can only be done if these groups of posters are controlled by very few individuals (“senders”). Indeed it is extremely unlikely (under one chance in a billion) that such correlations are due to chance.

Figure 4. Sychnometric estimates for a period of 800 seconds on a portion of the network



Thin peaks, indicated by red vertical lines, are separated by 2 minutes (120 seconds). Broader peaks, indicated by blue vertical lines, are separated by 3 minutes

Based on earlier estimates on latency distribution to Twitter’s CDN nodes, we can see that all the 2-minute peaks originate from a single sender (using several dozen accounts), and all the 3-minute peaks originate from two, slightly distant senders (again using many accounts). We shall call members of such collusions seeds.

From a network point of view, each seed is the centre of a user community, and typically two such communities are fairly distant. But the operators planting seeds across the network have tight control over the content and can efficiently give the impression that some information is true because diffusion patterns ensure a quick propagation. After a few minutes, several *a priori* independent sources convey this message.

The amount of precision required to operate all seeds is compatible with the use of dedicated computer programs (such wide-scale coordination at the sub-second level is very unlikely to be a manual labour); however the choice of placement and the contents are engineered strategically.

4. CONCLUSION AND FURTHER WORK

4.1. On the importance of being timely

This study shows how to leverage timing measurements to extract hitherto hidden information about a network’s operation, highlighting the structure of influence operations and calling for an integration of related measures in the representations of cyberspace. The extremely precise and abundant availability of time measurements makes it a reliable and efficient material.

4.2. Exploiting latent linguistic information

The same observations made for networks and network protocols hold for messages themselves, when they are exchanged between communities. To make the message understandable to its intended recipients, users have to convene of some format, which forces them to abide to some rules (grammatical or other); some users are less literate than others; messages have a given length, are typed at a certain speed, etc. In a sense, there is linguistic metadata readily available. In a network context, it is not possible to correctly participate in a communication without giving some reasonable accurate information, such as an IP address to establish a TCP connection; likewise participants in a conversation will leak some details about them or relations between them.

These two approaches –leveraging network and linguistic metadata– combine naturally and provide insights about how communities are formed; notions of network and linguistic centrality shed light on how a network operates and which elements hold influence over which others.

REFERENCES

Abdou A., Matrawy A., van Oorschot P. C., 2015, "Accurate One-Way Delay Estimation With Reduced Client Trustworthiness", *IEEE Communications Letters*, 19(5), p. 735-738.

Bacry E., Mastromatteo I., Muzy J.-F., 2015, "Hawkes processes in finance", *Market Microstructure and Liquidity*, 1(01), p. 1550005-1550064.

Douzet F., Desforges A., Limonier K., 2014, "Géopolitique du cyberespace : 'territoire', frontières et conflits", *Proceedings du 2^{e} colloque international du CIST*, Paris, CIST, p. 173-178.

Gummadi K.P., Saroiu S., Gribble S.D., 2002, "King: Estimating latency between arbitrary internet end hosts", *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurment*, p. 5-18.

Limonier K., 2014, "La Russie dans le cyberespace : représentations et enjeux", *Hérodote*, 2014/1, n° 152-153, p. 140-160.

Lombardi M. A., Nelson L. M., Novick A. N., Zhang V. S., 2001, "Time and frequency measurements using the global positioning system", *Cal Lab: International Journal of Metrology*, 8.3, p. 26-33.

Nygren E., Sitaraman R.K, Sun J., 2010, "The Akamai network: a platform for high-performance internet applications", *ACM SIGOPS Operating Systems Review*, 44(3), p. 2-19.

THE AUTHOR

Rémi Géraud
Département d'informatique de l'ENS, CNRS, PSL Research University
remi.geraud@ens.fr