Detection of Event Precursors in Social Networks: A Graphlet-Based Method

Hiba Abou Jamra, Marinette Savonnet, Eric Leclercq

Detection of event precursors in social networks: A graphlet-based method

Hiba Abou Jamra, Marinette Savonnet, and Éric Leclercq

Laboratoire d'Informatique de Bourgogne - EA 7534
Univ. Bourgogne Franche-Comté
Hiba_Abou-Jamra@etu.u-bourgogne.fr
Marinette.Savonnet@u-bourgogne.fr
Eric.Leclercq@u-bourgogne.fr
https://lib.u-bourgogne.fr

Abstract. The increasing availability of data from online social networks attracts researchers' interest, who seek to build algorithms and machine learning models to analyze users' interactions and behaviors. Different methods have been developed to detect remarkable precursors preceding events, using text mining and Machine Learning techniques on documents, or using network topology with graph patterns.
Our approach aims at analyzing social networks data, through a graphlets enumeration algorithm, to identify event precursors and to study their contribution to the event. We test the proposed method on two different types of social network data sets: real-world events (Lubrizol fire, EU law discussion), and general events (Facebook and MathOverflow). We also contextualize the results by studying the position (orbit) of important nodes in the graphlets, which are assumed as event precursors. After analysis of the results, we show that some graphlets can be considered precursors of events.

Keywords: Graphlets · Event Precursors · Social Networks

1 Introduction

Online social networks (OSN) play an essential role in individuals' and businesses' daily lives. Due to social interactions between individuals in these networks, scientists have an opportunity to observe and analyze increasing amounts of data to extract value and knowledge.

Disease outbreaks, environmental and industrial crises present challenges to researchers in different domains such as economy, finance, earth sciences, epidemiology, and information science. Detection of weak signals can be a key for anticipating changes in advance and avoid letting them cause surprise [10]. OSN enhance the emergence of echo chambers where ideas are amplified and can conduct to a digital crisis. To limit negative publicity (known as "bad buzz"), organizations should be vigilant to weak signals. Detection of significant patterns

or motifs helps to understand the network dynamics and identify or predict complicated situations. Network topological properties such as density, assortativity, and degree centrality help to understand the network's global structure.

This article introduces an approach to help experts detect weak signals by topological analysis of the Twitter network. Our main contributions are 1) identification of graphlets as event precursors; 2) evaluation of the identified graphlets about their participation in the event; 3) contextualization of the results to help experts in interpretation; 4) evaluation of the proposed method using existing real data sets obtained from the Cocktail project and well-known data sets used as a benchmark. Cocktail is an interdisciplinary project aiming to develop a platform that will enable organizations to build a communication strategy, anticipate a crisis via a communication response, and adapt their industrial offers.

The rest of this article is organized as follows: Section 2 introduces some background on weak signals, event precursors, and describes similar works. In section 3, after a brief reminder on graphlets concept, we explain and illustrate the proposed method starting from time series of social networks data to event precursors identification, and the study of the correlation between precursor graphlets and the event of interest. Section 4 introduces the experimental part: it describes the main characteristics of the used data sets. In section 5, we test the proposed approach on real events based on industrial and environmental crises, along with experiments on benchmark network models to evaluate and verify this approach. Finally, conclusions and future perspectives are presented in section 6.

2 Related Work

In a digital society, detection of weak signals has become necessary for decision-makers in industrial and commercial policy and communication strategy while projecting future scenarios. Weak signals can be the precursors of future events. The detection of these signals can either transform them towards a trend or an event in the future or stop their evolution for controlling and preventing future crises. Ansoff [2] was the first to propose the concept of a weak signal for strategic planning through environmental analysis. He defines weak signals as the first symptoms of strategic discontinuities that act as early warning information of low intensity, which can be the initiator of an important trend or event. Table 1 presents terms and definitions qualifying weak signals by social scientists. Event precursors and weak signals are two concepts with strong proximity. Generally speaking, a precursor is in a relationship with the event of interest. It is any behavior, situation, or group of events that is a leading indicator of future incidents or consequential events [6]. In the following, we present several studies related to our work. We can classify these studies into three categories: 1) text mining and Natural Language Processing (NLP); 2) Machine Learning (ML) for identification and forecasting; and 3) motifs or patterns.

Many text mining and NLP approaches have been proposed, where Web documents are analyzed through a quantitative analysis of keywords. Yoon et

Table 1. Weak signals definitions and terms

Source	Definitions and terms
Ansoff 1975 [2]	Incomplete information, imprecise, fragmentary
Godet 1994 [8]	Low intensity, low visibility
Coffman 1997 [5]	Initiator of an important event, of a future trend
Hiltunen 2010 [10]	Low utility, meaningless when analyzed individually
Welz 2012 [23]	But can make sense if seen as a set of information

al. [24] have proposed two indicators: the degree of visibility based on keyword frequency and the degree of diffusion based on document frequency and considering their rates of increase in time. A keyword that has low visibility and a low diffusion level is considered a weak signal. Other studies leaned on these two indicators by adding a context to a list of keywords and used, for example, topic modeling such as LDA (Latent Dirichlet Allocation) [14, 15] and clustering algorithms such as k-Means or k-Medoids [16].

Ning et al. [17] developed a model of multiple instance learning algorithms, based on supervised learning techniques, to formulate the precursor identification and the forecasting issue. The model consists of assigning a probability to collected news articles associated with targeted events (protests in their study). The greater probability is, the more the news article is considered as a precursor containing information about this event's cause. Another study by Ackley et al. [1] adopted supervised learning techniques (Random Forest and Sequential Backward Selection algorithms) in the commercial aviation operation domain to analyze and track critical parameters leading to safety events in the approach and landing phases.

Furthermore, some researchers were interested in identifying specific patterns in networks, known as motifs, which could be considered as event precursors. Baiesi et al. [3] presented a method that studies correlations within graphs of upcoming earthquakes using tools of network theory. They measured the distance between network nodes along with the clustering coefficient, which reflected intentionally basic mechanisms of seismic movements and earthquake formation/propagation. After applying statistical tools on the network topology, they found that simple motifs such as special triangles constitute an interesting type of precursors for significant events. Later on, several approaches studied the identification and the role of motifs in critical events such as crime analysis [7] and ongoing attacks detection [13].

These works aimed to identify weak signals relying on text mining techniques and network theory tools. The last one leads us to our hypothesis that graphlets, which are particular motifs, can be precursors of events. But to the best of our knowledge, there has not been a graphlet-based solution to detect event precursors in social networks and assess their relationship with the event of interest. We present our proposed approach in the upcoming section.

3 Graphlets as potential event precursors

The most known characteristics of weak signals are usually hard to quantify, so we prefer to rely on the notion of event precursors of small intensity to obtain a more precise definition, by considering an event as an activity peak and a precursor as a signal of lower importance or intensity, being in a correlation with the event.

Conventional methods based on simple statistical techniques are not able to identify event precursors easily. Instead, they are helpful to identify events such as the family of ARIMA, EDM, HDC algorithms [20]. We choose to explore another approach based on the assumption that networks' topology plays an essential role in information propagation, hence in the formation of an event, so we assume that graphlets found in social networks can be considered as potential event precursors, just as cliques are for communities. They have proven their worth in numerous contexts in network research [12].

In this section, we investigate the following questions: Can graphlets be identified as event precursors? Can these precursors be qualified as weak signals prior to the event of interest? Before going into details, we present the essential notion of graphlets. We describe how to prepare and transform data to enumerate graphlets in a temporal graph built from interactions between users and discover the potential event precursors' graphlets.

3.1 Graphlets in a nutshell

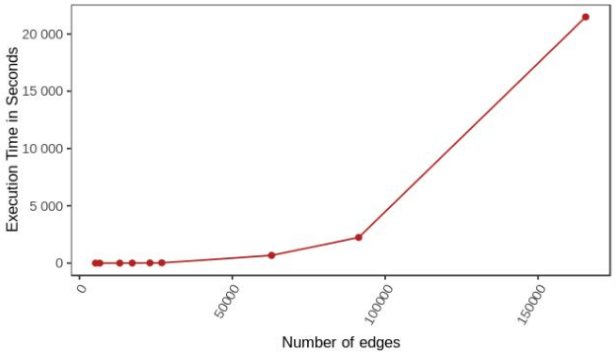
Graphlets, first introduced by Pržulj [19], are particular types of motifs in a network, and thanks to their predefined sizes and shapes, they are easy to interpret by experts in the domain, such as social scientists or political scientists. A graphlet is a connected induced non-isomorphic subgraph (2 to 5 nodes) chosen from the nodes of a large graph. 30 graphlets from G_0 to G_{29} with up to 5 nodes are possible: the G_0 graphlet of size 2, two graphlets of size 3 which are G_1 and G_2  , 6 graphlets of size 4, and 21 graphlets of size 5. Orbits, or positions, represent the equivalence classes of graphlets [18]. They are the positions to which nodes belong in the 30 graphlets; nodes belonging to the same orbit are interchangeable. For example, the star-shaped G_4 graphlet  consists of two positions; one of them is central (orbit 7) occupied by one node, and the other is peripheral (orbit 6) and shared between the remaining three nodes that are interchangeable.

There exist several algorithms to enumerate graphlets and orbits of a graph. A survey was made by Ribeiro et al. in 2019 [21], in which they provided an overview of the existing algorithms for subgraph counting, classified these algorithms, and highlighted their main advantages and limitations. They explored the methods for counting subgraphs from three perspectives: 1) exact counting algorithms (*e.g., ESU/FANMOD, RAGE, Orca*); 2) approximate counting algorithms (*e.g., ESA, RAND-ESU*); 3) parallel processing algorithms (*e.g., DM-ESU, GPU-Orca*). The survey provides valuable insight from a practical point

of view of the algorithms and their existing implementations with a trade-off between accuracy and execution time.

To choose the most convenient algorithm for counting graphlets and orbits in the studied graph structures, we have defined 3 essential criteria: 1) exact counting of graphlets that are up to five nodes, to maintain the interpretability of the results; 2) orbits counting for the study of nodes positions within each graphlet; 3) availability of source code. We rely on the Orca algorithm proposed by Hočevar and Demšar in 2014 [11], which is an exact counting algorithm, coming from an analytic approach based on matrix representation, and works by setting up a system of linear equations per node of the input graph that relate different orbit frequencies. It counts small subgraphs up to 5 nodes and focuses on orbits counting. Considering e as the number of edges and d the maximum degree of nodes, its time complexity is of $\mathcal{O}(ed)$ for four-node graphlets and $\mathcal{O}(ed^2)$ for five-node graphlets. We performed an experimental analysis to evaluate Orca's implementation complexity. With up to 15 000 edges in a graph, the calculation time is less than 5 seconds, but it reaches 6 hours with up to 160 000 edges. Figure 1 shows execution time based on the number of edges.

Fig. 1. Experimental evaluation of Orca's complexity

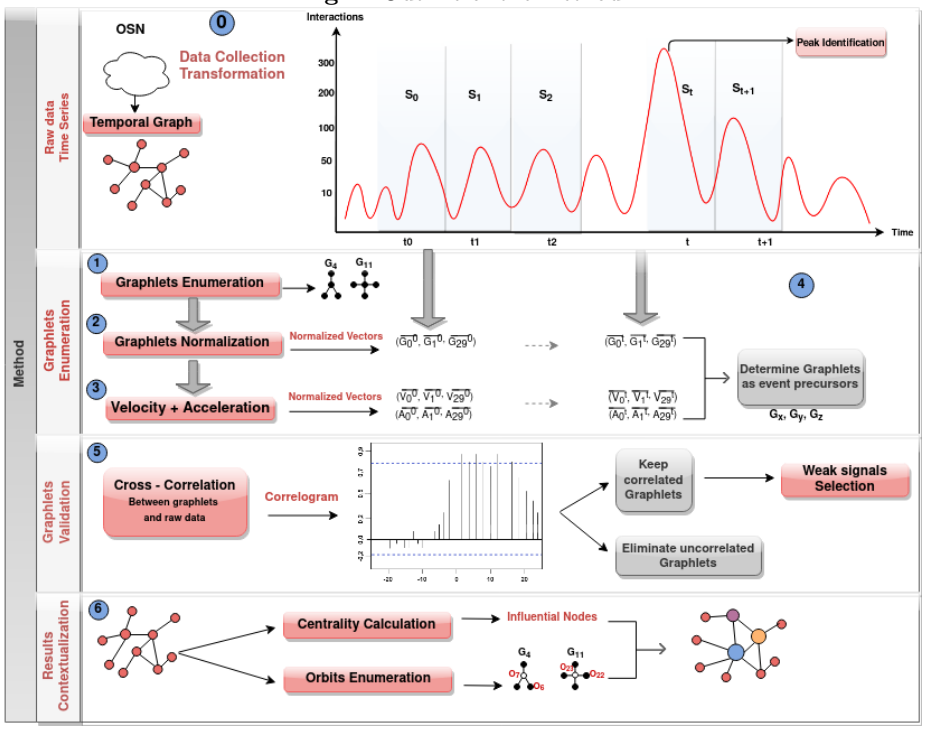


3.2 Proposed method

We propose a graphlet-based analysis method that facilitates the results' interpretation. Once the potential precursors have been revealed, it is still necessary to validate the fact that they are weak signals, determine their link with the studied event, and then allow experts to understand their role. We prove experimentally that these graphlets can be event precursors. We present our method consisting of six steps, depicted in figure 2.

0. The first step is to build a time series from social networks data. Once raw data is collected, for example, tweets in JSON format, some interactions

Fig. 2. Outline of the method



of interest are selected (e.g., retweet, quote, mention), and a graph structure is generated as a tuple with three components representing interactions between entities at a given date (e.g., `(user1,user2,124354432)`). A time series X is created from the number of interactions selected between all pairs of nodes. It is a sequence of n elements $X = (x_i)_{1 \leq i \leq n} = (x_1, x_2, \ldots, x_n)$. A method to remove the seasonal part from the original time series is then applied [4]. We consider an event as an activity peak resulting from a variation in the interactions between entities, and the peak is identified either manually or by event detection algorithms. Before and during the event, the time series is divided into snapshots S^t according to the duration or importance of the event (e.g., a day, 12 hours, 6 hours, 1 hour), in a way to have sub time series of the original series: $S^t = (x_i)_{t \leq i < t+d}$ with d the constant duration of a snapshot, and the constraint that all the S^t form a partition of the original series X.

1. Enumeration of graphlets for each snapshot determines a topological signature before and during the event. Snapshots S^t are represented as components of a numerical vector $(G_0^t, G_1^t, \ldots, G_{29}^t)$, G_x^t is the number of graphlets

of type x in the snapshot S^t. We rely on the Orca algorithm[1], it provides an acceptable runtime as all snapshots contain at most a few thousand edges (see figure 1).

2. We apply a normalization procedure on these vectors to re-scale their values to a particular magnitude for further measuring and calculations. This step is of significant importance as it should not hide small signals but instead make them comparable to others. The procedure relies on a framework proposed by D.Goldin and P.Kanellakis [9] in which they study the similarity between two queries relating to a temporal database. Two real numbers a and b define a transformation $T_{a,b}$ on X by joining each x_i with $a \times x_i + b$.
$\overline{X}$ represents the normal form of X calculated by:

$$\overline{X} = T^{-1}_{\sigma,\mu}(X) = T_{\frac{1}{\sigma},-\frac{\mu}{\sigma}}(X)$$

in which $\mu(\overline{X}) = 0$ and $\sigma(\overline{X}) = 1$, μ is the mean and σ the standard deviation. Therefore, the mean of each graphlet type G_x for all snapshots is calculated as:

$$\mu(G_x) = \frac{1}{s}\sum_{t=1}^{s}(G_x^t) \quad \forall x \in \{0, \ldots, 29\}, s \text{ is the number of snapshots}$$

Then the standard deviation is calculated as:

$$\sigma(G_x) = \sqrt{\frac{\sum_{t=1}^{s}(G_x^t - \mu(G_x))^2}{s-1}} \quad \forall x \in \{0, \ldots, 29\}$$

By applying this normalization procedure for each of the snapshots S^t, each component of its vector G_x^t is normalized by:

$$\overline{G_x^t} = \frac{(G_x^t) - \mu(G_x)}{\sigma(G_x)}$$

3. From the normalized values obtained, the evolution of all the vector components is studied via the calculation of their velocity and acceleration, with the purpose to highlight the graphlets that come out quickly before the other types. The calculation of these attributes is as follows:
   - Velocity: $\overline{V_x^t} = \overline{G_x^{t+1}} - \overline{G_x^t} \quad \forall x \in \{0, \ldots 29\}$
   - Acceleration:

$$\overline{A_x^t} = \frac{\Delta V_x}{\Delta t} = \overline{V_x^{t+1}} - \overline{V_x^t} \quad \forall x \in \{0, \ldots 29\}, \Delta t = 1 \text{ between snapshots}$$

4. We observe the obtained results in steps 2 and 3 to capture significant variations in their values before the activity peak. We choose the k graphlets with the highest velocity and acceleration values as potential precursors of events.

[1] https://rdrr.io/github/alan-turing-institute/network-comparison/src/R/orca_interface.R

5. This step aims to validate the potential precursors' graphlets by eliminating those irrelevant (false positives) and maintaining the pertinent ones supposed as weak signals (true positives). Although keeping some false positives can help social scientists to examine the information behind critical situations. It is composed of two stages: 1) we evaluate cross-correlation between each precursors' graphlet time series and the original interactions time series, and 2) for correlated graphlets, we quantify their contribution to the global evolution of graphlets to confirm if they are weak signals or not.

   Cross-correlation [2] is used to validate the intrinsic properties of the method. It is a linear measure of similarities between two time series X and Y, which helps evaluate the relationship between two series over time [22]. An offset/lag h is associated with this measure, knowing that if $h < 0$ then X could predict Y, and if $h > 0$ then Y could predict X.

   Weak signals selection is a simple ratio calculation that measures the correlated graphlets' contribution to the global evolution of graphlets for the studied period. From the correlated graphlets found, the total number of a graphlet type x in all snapshots is divided by the total number of graphlets for all snapshots, as follows:

$$R(G_x) = \frac{\sum_{t=1}^{s}(G_x^t)}{T(G)}$$

   and $T(G) = \sum_{t=1}^{s}(\sum_{x=0}^{29}(G_x^t))$, with s the number of snapshots.
   The resulted ratios R, are sorted in ascending order, to verify if the identified correlated graphlets remain at the top of the list; if so, they are qualified as weak signals, the other graphlets are eliminated.

6. This step aims to provide adequate analysis elements to domain experts to interpret the previous steps' obtained results and respond to potentially critical situations. For each orbit (i.e. the position, or the node's role in the graphlet) of graphlets considered as weak signals, we count how many times nodes of the initial graph appear in these graphlets. To restrict the information to study and facilitate the interpretation, we consider only the most influential nodes, hence the PageRank algorithm is used to help to identify these nodes in the graph.

4 Data description

We describe in this section the data sets used for our experiments. To this end, the selected data sets include a sequence of temporal interactions between users. Two first data sets represent real case scenarios, and the other two sets[3] are social benchmark networks used to confirm our method.

[2] implemented with the R package *tseries*: https://www.rdocumentation.org/packages/tseries/versions/0.1-2/topics/ccf

[3] https://snap.stanford.edu/data/#socnets

Twitter - Lubrizol fire: This network contains tweets published after a fire broke out at the Lubrizol factory in Rouen-France. From the raw Twitter data, the corpus contains tweets between midnight of October 28, 2019, and midnight of October 30, 2019. The reduced corpus consists of 18,914 tweets, 12,187 of these tweets are original, and 1,984 include mentions which are the interaction type studied in this example.

Twitter - European CAP Law: This dataset contains tweets published in conjunction with the European Council meeting held on October 20, 2020, that lead to the announcement of the Common Agricultural Policy Law (CAP). The dataset comprises tweets collected from midnight of October 17, 2020, to midnight of October 20. It consists of 4,679 tweets, from which 807 are original, and 3,872 include mentions and retweets, which are the interaction type we study in this experiment.

MathOverflow Network: This network contains temporal user interactions from the Stack-Exchange site "Math-Overflow" consisting of three interaction types: 1) answering a question; 2) commenting on another user's question; 3) commenting on another user's answer. The used data set is extracted from the original sample, and consists of 1,400 relations from October 27, 2010, till October 30, 2010.

Facebook Network: This is a network representing a subset of posts to other user's walls on Facebook. The raw data sample is collected from October 2004 to January 2009, and we minimize the set to include the detailed relations between 05 January 2009, and 07 January 2009, consisting of up to 8,790 interactions between users.

5 Experiments, results and discussion

In this section, we present experiments that aim to validate the proposed method by detecting graphlets that are supposed to be weak signals, supplemented by contextualization elements so that experts can trigger actions. We apply experiments on four different data sets, two of them are the subject of critical situations in industry and agriculture, and the remaining ones belong to random events. The Cocktail platform collected the first two data sets, domain experts provided the accounts and keywords needed for the collection. The two other data sets were used to validate the approach.

5.1 Industrial crisis: Twitter - Lubrizol fire

The first experiment of the proposed approach was carried out on the Twitter Lubrizol network. Our event of interest is the unexpected visit of President Macron to Rouen, October 30, 2019, around 6 p.m. Therefore, the study period is reduced to two days before the event (28 and 29), along with the event's day (30). After step 0, we obtain a temporal graph that contains 2,231 nodes and 3,821 edges[4].

[4] The difference between the number of tweets in section 4 and the number of nodes and edges is since several tweets can produce the same interaction.

We choose to work with snapshots of one hour, to capture the biggest number of graphlet types, especially those with complex shapes, which will help with a finer interpretation of the results. The graphlets number is calculated for each snapshot and the resulting values are normalized. Next, velocity and acceleration are measured for the normalized graphlet values. After analysis of the three computed attributes, we notice an increase in certain graphlets' number and velocity on October 30 starting at 4 p.m., like G_2 , G_5  , G_8  and G_{27}  . Therefore, we consider these graphlet types as potential event precursors. Table 2 presents graphlets number, velocity, and acceleration results for certain graphlet types, for the snapshots corresponding to three hours before the event. It compares the evolution of the attributes mentioned above between the graphlets that evolved starting at 4 p.m. (supposed precursors), and other graphlets that did not show remarkable variations for the same snapshots. The notable changes in attributes' values are highlighted in blue. We notice that other graphlet types like G_4  and G_{11}  , start increasing from 6 p.m, which is the snapshot of the activity peak, and hence they are aligned with the event.

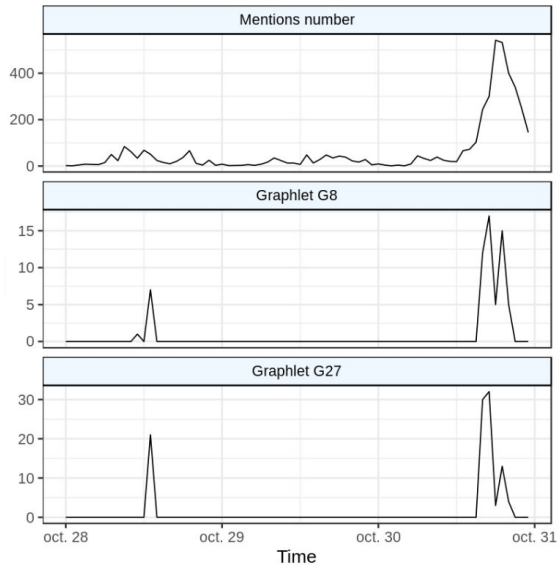
Table 2. Enumeration results of some graphlets before the event. The highlighted values correspond to the potential precursor's graphlets

	S^t: 30/10 3p.m-4p.m			S^{t+1}: 30/10 4p.m-5p.m			S^{t+2}: 30/10 5p.m-6p.m		
Graphlet	$\overline{G_x^t}$	$\overline{V_x^t}$	$\overline{A_x^t}$	$\overline{G_x^t}$	$\overline{V_x^t}$	$\overline{A_x^t}$	$\overline{G_x^t}$	$\overline{V_x^t}$	$\overline{A_x^t}$
G2	-0,1592	0,1869	0,3738	3,0657	3,2248	3,0379	3,4863	0,4206	-2,8042
G5	-0,1881	0,1417	0,1102	2,9505	3,1387	2,9970	3,7116	0,7610	-2,3776
G8	-0,2796	0	0	3,4544	3,7340	3,7340	5,0102	1,5558	-2,1781
G27	-0,2364	0	0	5,2868	5,5233	5,5233	4,3715	-0,9152	-6,4385
G17	-0,1817	0	0,0012	0,0212	0,2030	0,2030	0,5591	0,5379	0,3348
G22	-0,1623	0	0,0006	0,0355	0,1979	0,1979	0,1083	0,0727	-0,1252

Next, we validate the potential precursors and select the ones supposed to be weak signals. We apply cross-correlation between the initial time series and the ones belonging to precursor graphlets. The time series of G_2, G_5, G_8 and G_{27} present correlations with a positive lag h of one and two hours with the initial time series, having significant values equal to 0.8, which indicates that the number of interactions in the initial series follows with a lag of 1 or 2 hours the number of graphlets. The calculated ratios highlight the weak presence of these correlated graphlets in the rise of mentions number, compared with other strong graphlets like G_{11}, hence G_2, G_5, G_8 and G_{27} are considered weak signals.

Figure 3 represents some of the considered weak signals' time series, compared to the initial mentions time series.

Fig. 3. Initial mentions time series vs. G_8 and G_{27} graphlets time series, considered weak signals



A fine-grained experiment is carried out to contextualize the obtained results in the previous steps: we calculate the number of times an influential node is in an orbit of a selected graphlet. We find users like `manon_leterq` and `massinfabien` journalists, and `76actu` the local information site, having a rise in the number of orbits of the selected graphlets, starting at 4 p.m (two hours before the event). Table 3 presents an extract of the number of times the above influential users appear in the orbits of graphlets G_2 (O_3) and G_{27} (O_{68} and O_{69}). The remarkable increase in values is highlighted in blue. We did the same calculations with a user-chosen randomly `OTT_44380`; the results show that he appears little in the graphlets.

We repeat the same experiment on different time windows. In the 6-hours snapshot, we were able to extract the same graphlets; on the contrary, we could not confirm the exact time of their appearance due to the window's large size. A finer study on 30 and 15 minutes snapshots (containing fewer edges) led to a partial vanishing of complex graphlets like G_8 and G_{27} over time. The absence of these complex graphlets results in information loss and makes decision-making more difficult. We rely on providing enough information to the experts to take preventive actions.

Table 3. Extract of influential users and their orbits enumeration results for some of the precursor graphlets

	S^t: 30/10 3p.m-4p.m			S^{t+1}: 30/10 4p.m-5p.m			S^{t+2}: 30/10 5p.m-6p.m		
User	O3	O68	O69	O3	O68	O69	O3	O68	O69
OTT_44380	0	0	0	0	0	0	7	18	3
manon_leterq	0	0	0	12	72	0	20	68	0
76actu	0	0	0	31	18	49	30	5	70
massinfabien	0	0	0	10	63	0	6	32	0

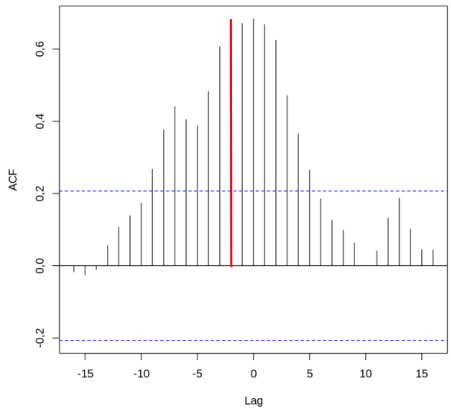
5.2 Environmental crisis: Twitter - CAP Law

In this experiment, we are interested in the mentions and retweets published after European Council meetings held in late October 2020 for negotiation on the post-2020 Common Agricultural Policy (CAP) reform package, which later initiated an agreement on the proposed CAP project. The event corresponds to the 20th of October at noon, where the European Council took a position towards the CAP project. Thus, we focus on the two days preceding the event (18 and 19) and the day of the event (20). The initial graph contains 2,535 nodes and 7,897 edges. The corresponding time series of the interactions is created and divided into snapshots of one hour each. The enumeration of graphlets and the calculation of velocity and acceleration allow us to extract the most pertinent graphlets G_{15}  , G_{18}  , G_{21}  , and G_{28}  as potential event precursors, due to the rise of their values between one and three hours before the event. Other graphlet types such as G_9 arise in parallel with the event.

The cross-correlation applied to precursor graphlets shows on the one hand that G_{15} and G_{21} present a positive correlation of two and three hours lags, respectively, with the initial time series, having values equal to 0.7 and 0.6 respectively (see Figure 4). G_{18} also shows a positive correlation of one hour lag with a value equal to 0.7. On the other hand, G_{28} did not present a positive correlation with the initial time series, hence it is not a weak signal. Furthermore, ratios are calculated for the correlated graphlets. These graphlets have low ratios compared with other graphlets types, so they are considered weak signals.

In the last step, orbits in graphlets are enumerated for the identified influential users. The results show that, for instance, a user like `pcanfin` (Chair of the environment committee of the European Parliament) appears for the first time at 10 a.m of the event day, in orbit O_{51} of graphlet G_{21}  , and at 11 a.m in orbit O_{34} of graphlet G_{15}  . Another user, `TheProgressives` (representing Socialists and Democrats Group in the European Parliament), appears at 10 a.m in the orbits of G_{15}, G_{18} and G_{21}, but shows up strongly at 11 a.m in orbits O_{34} and O_{51}. Then the number of orbits of these users starts decreasing towards the event. These users interacted against the law a few hours before the announcement of the council's decision, and their positions in these graphlets (closed connected structure) above can reveal their role in a strongly connected

Fig. 4. Correlation of G_{15} with the initial mentions-retweet time series



community of users that might share the same political opinion in terms of reactions to the ongoing situation.

5.3 Random Events: MathOverflow and Facebook

The objective of these two experiments is to verify and validate the proposed method in terms of reproducibility and results interpretation. Our method is applied to the benchmark network models MathOverflow and Facebook interactions. The event is unknown here, so we select a peak activity in the corresponding time series of each network and consider it as the event of study, to evaluate the previously obtained results for Twitter network data. We also work by snapshots of one hour.

The peak selected from the MathOverflow time series (the graph contains 414 nodes and 966 edges) belongs to activity on October 29, 2010, at 11 p.m. Enumeration results show remarkable variations in certain graphlets numbers on October 29 starting at 10 p.m. We find graphlets G_3 , G_9 and G_{10} increasing first follows them the G_2 at 11 p.m., time of the peak activity. After that time, the calculated numbers start decreasing accordingly. The correlation study was not able to find positive correlations between the time series of these graphlets and the initial time series before the peak. Moreover, the enumeration of orbits in the last step was not entirely relevant since the nodes belong to anonymous users, hence the results cannot be reflected into real scenarios for analysis.

Furthermore, the Facebook data set was studied by snapshots of one hour to ease the discovery of significant precursors before the selected peak activity in the related time series (the graph contains 6,726 nodes and 6,677 edges). The peak corresponds to an event on January 07, 2009, starting at 6 a.m. We notice a prominent rise in numbers of G_2 and G_6 in the evening before the peak activity,

at 6 p.m. Just as the Mathoverflow dataset, these identified graphlets can not be considered as event precursors as they did not show significant correlation results with the initial time series.

The obtained results in these benchmark networks lead us to the interpretation synthesis that these are too generalist data sets, and we can track no targeted event in the real world. Moreover, we could not identify weak signals, since most of the graphlets participate strongly in the rise of interactions between users. Data and experimental programs are available under https://github.com/hibaaboujamra/EventPrecursorsGraphlets.

6 Conclusion and perspectives

We have studied the hypothesis of discovering whether the graphlets are precursors for occurring events, and developed a method to evaluate and confirm this hypothesis. The proposed approach allows identifying graphlets as precursors of events and targeting those that constitute weak signals.

We performed quantitative and qualitative analyses using graph enumeration and correlation measures. The experimental results confirm that our method was able to identify event precursors and target those that can be weak signals two hours before an event.

Moreover, the last step of contextualization provides rich elements to domain experts for further analysis and interpretation of the results to react accordingly in case of critical situations.

In future works, we want to extend the experiments to other types of networks as the hashtags co-occurrence, for example, and larger networks, and automate all the method's steps. We observed experimentally that graphlets are good precursors of events, hence we attach currently to establish proof of causality between these graphlets and the event, through statistical methods like the causal inference or the Granger causality. Further investigations will consider iterating our method to eliminate nodes continuously from graphs to decrease certain graphlets' predominance and allocate the space to discover other graphlet types as event precursors to obtain a hierarchical graphlet decomposition.

Acknowledgments

This work is supported by the program "Investissements d'Avenir", ISITE-BFC project (ANR contract 15-IDEX-0003), https://projet-cocktail.fr/

References

1. Jamey L Ackley, Tejas G Puranik, and Dimitri Mavris. A Supervised Learning Approach for Safety Event Precursor Identification in Commercial Aviation. In *AIAA Aviation Forum*, page 2880, 2020.
2. H Igor Ansoff. Managing strategic surprise by response to weak signals. *California management review*, 18(2):21–33, 1975.

3. Marco Baiesi. Scaling and precursor motifs in earthquake networks. *Physica A: statistical mechanics and its applications*, 360(2):534–542, 2006.
4. Jason Brownlee. *Introduction to time series forecasting with python: how to prepare data and develop models to predict the future*. Machine Learning Mastery, 2017.
5. Brian Coffman. Weak signal research, part I: Introduction. *Journal of Transition Management*, 2(1), 1997.
6. William R Corcoran. Defining and analyzing precursors. In *Accident precursor analysis and management: Reducing technological risk through diligence*, pages 79–88. National Academy Press Washington, DC, 2004.
7. Toby Davies and Elio Marchione. Event networks and the identification of crime pattern motifs. *PloS one*, 10(11):e0143638, 2015.
8. Michel Godet. *From anticipation to action: a handbook of strategic prospective*. UNESCO publishing, 1994.
9. Dina Q Goldin and Paris C Kanellakis. On similarity queries for time-series data: constraint specification and implementation. In *International Conference on Principles and Practice of Constraint Programming*, pages 137–153. Springer, 1995.
10. Elina Hiltunen. Weak Signals in Organisational Futures. *Aalto University School of Economics, Aalto*, 2010.
11. Tomaž Hočevar and Janez Demšar. A combinatorial approach to graphlet counting. *Bioinformatics*, 30(4):559–565, 12 2014.
12. Yuriy Hulovatyy, Huili Chen, and Tijana Milenković. Exploring the structure and function of temporal networks with dynamic graphlets. *Bioinformatics*, 31(12):i171–i180, 2015.
13. Krzysztof Juszczyszyn and Grzegorz Kołaczek. Motif-based attack detection in network communication graphs. In *IFIP International Conference on Communications and Multimedia Security*, pages 206–213. Springer, 2011.
14. Krigsholm and Riekkinen. Applying Text Mining for Identifying Future Signals of Land Administration. *Land*, 8(12):181, Nov 2019.
15. Julien Maitre, Michel Ménard, Guillaume Chiron, Alain Bouju, and Nicolas Sidère. A Meaningful Information Extraction System for Interactive Analysis of Documents. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 92–99. IEEE, 2019.
16. Antonio Leonardo Martins Moreira, Thomas Wiliam Norio Hayashi, Guilherme Palermo Coelho, and Ana Estela Antunes da Silva. A Clustering Method for Weak Signals to Support Anticipative Intelligence. *International Journal of Artificial Intelligence and Expert Systems (IJAE)*, 6(1):1–14, 2015.
17. Yue Ning, Sathappan Muthiah, Huzefa Rangwala, and Naren Ramakrishnan. Modeling precursors for event forecasting via nested multi-instance learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1095–1104, 2016.
18. Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 01 2007.
19. Nataša Pržulj, Derek G. Corneil, and Igor Jurisica. Modeling interactome: scale-free or geometric? *Bioinform.*, 20(18):3508–3515, 2004.
20. Soumi Ray, Dustin S McEvoy, Skye Aaron, Thu-Trang Hickman, and Adam Wright. Using statistical anomaly detection models to find clinical decision support malfunctions. *Journal of the American Medical Informatics Association*, 25(7):862–871, 2018.
21. Pedro Ribeiro, Pedro Paredes, Miguel EP Silva, David Aparicio, and Fernando Silva. A survey on subgraph counting: concepts, algorithms and applications to network motifs and graphlets. *arXiv preprint arXiv:1910.13011*, 2019.

22. Brian D Ripley and WN Venables. *Modern applied statistics with S*. springer, 2002.
23. Kirill Welz, Leo Brecht, Anja Pengl, Julian V Kauffeldt, and Daniel RA Schallmo. Weak signals detection: Criteria for social media monitoring tools. In *ISPIM Innovation Symposium*, page 1. The International Society for Professional Innovation Management (ISPIM), 2012.
24. Janghyeok Yoon. Detecting weak signals for long-term business opportunities using text mining of web news. *Expert Systems with Applications*, 39(16):12543–12550, 2012.