

Détection de précurseurs d'évènements basés sur les motifs dans les réseaux sociaux

Hiba Abou Jamra*, Marinette Savonnet*, Éric Leclercq*

*Laboratoire d'Informatique de Bourgogne - EA 7534
Univ. Bourgogne Franche-Comté
9, Avenue Alain Savary, F-21078 Dijon - France
Hiba_Abou-Jamra@etu.u-bourgogne.fr
Marinette.Savonnet@u-bourgogne.fr
Eric.Leclercq@u-bourgogne.fr

Résumé. Les données issues des réseaux sociaux suscitent l'intérêt des chercheurs qui développent des algorithmes et des modèles d'apprentissage automatique pour analyser les interactions et les comportements des utilisateurs. Ces méthodes s'appuient sur la topologie du réseau pour représenter les changements structurels et pour détecter des précurseurs remarquables précédant généralement des évènements majeurs. L'étude présentée dans cet article vise à étudier si certains graphlets (motifs spécifiques) peuvent être considérés comme des précurseurs d'évènements. Nous expérimentons la méthode proposée sur trois ensembles de données de réseaux sociaux. Nous étudions également le rôle joué dans les graphlets (orbites) par les nœuds ayant une position centrale dans le graphe global. Après analyse des résultats, nous montrons que les graphlets constituent des précurseurs d'évènements à considérer.

1 Introduction

Les réseaux sociaux jouent un rôle important dans la vie quotidienne des particuliers et des entreprises. En raison des interactions sociales entre les individus à travers ces réseaux, les chercheurs ont la possibilité d'observer et d'analyser de grandes quantités de données afin d'extraire des connaissances.

Développer des méthodes pour détecter des évènements survenant dans les réseaux sociaux le plus tôt possible, les comprendre et les expliquer, est un défi de recherche important. L'identification des précurseurs d'évènements permet de déclencher des mesures préventives pour contrôler une épidémie ou pour étudier un phénomène. Ces méthodes peuvent être utilisées dans de nombreux domaines tels que l'économie, la finance, le marketing, les sciences de la terre, l'épidémiologie, le contrôle des « fake news », etc. Les précurseurs d'évènements sont également appelés signaux faibles en sciences sociales où la première définition a été élaborée dans le contexte de l'étude des processus de gestion stratégiques (Ansoff, 1975), mais aussi en économie et psychologie (Schoemaker et Day, 2009; Harrysson et al., 2014). Ces termes introduisent une notion d'émergence, signifiant que les précurseurs ou les signaux faibles ont une relation causale avec les évènements.

Les propriétés topologiques du réseau telles que la densité, l'assortativité et le degré de centralité servent à comprendre la structure globale du réseau, mais la détection de patterns significatifs est une autre étape pour comprendre la dynamique du réseau et identifier ou prévoir des situations problématiques. Nous formulons le problème d'identification des précurseurs d'évènements en utilisant le concept de graphlets. Nos principales contributions sont : 1) l'identification de graphlets comme précurseurs d'évènements ; 2) l'évaluation des graphlets identifiés et l'étude de leur émergence causale ; 3) la définition, la réalisation et l'interprétation d'expériences sur plusieurs réseaux temporels, dont un issu du projet COCKTAIL¹, et deux autres utilisés comme benchmark (Leskovec et Krevl, 2014).

Le reste de cet article est organisé comme suit : la section 2 présente le contexte général sur les précurseurs d'évènements et les signaux faibles, et décrit également des travaux similaires. Dans la section 3, après un rappel rapide du concept de graphlet et des algorithmes d'énumération des graphlets, nous expliquons et nous illustrons la méthode proposée à partir d'un graphe temporel jusqu'à la modélisation des graphlets, ensuite nous étudions la corrélation entre les différents indicateurs calculés. La section 4 introduit la partie expérimentale : elle décrit les propriétés topologiques des ensembles de données utilisés. Une étude de cas sur un évènement réel ainsi que des expériences sur deux réseaux benchmark pour confirmer les résultats obtenus sont présentées dans la section 5. Enfin, les conclusions et perspectives futures sont présentées dans la section 6.

2 Travaux connexes

Les précurseurs d'évènements et les signaux faibles sont deux concepts proches qui ont émergé de domaines différents. De manière générale, un précurseur est le facteur causal lié à un évènement important. Il s'agit de tout comportement, toute situation ou groupe d'évènements qui est un indicateur d'incidents futurs ou d'évènements consécutifs². La détection des signaux faibles est un enjeu important puisqu'elle permet d'anticiper des prises de décision en matière de politique industrielle et commerciale et de stratégie de communication tout en projetant des scénarios sur l'avenir. Les signaux faibles peuvent être les précurseurs d'évènements futurs. La première théorisation des signaux faibles a été proposée par Ansoff (1975) où il définit les signaux faibles comme les premiers symptômes de discontinuités stratégiques qui agissent comme une information d'alerte précoce, de faible intensité, pouvant être annonciatrice d'une tendance ou d'un évènement important. Par conséquent, l'identification des signaux faibles dans les données massives nécessite des méthodes d'analyse quantitative, souvent appelées « smart data », pour anticiper certains évènements. Lesca et Blanco (2002) ont proposé les caractéristiques suivantes pour identifier un signal faible : fragmentaire, peu visible, peu fréquent, utilité et fiabilité faibles. Néanmoins, ces caractéristiques sont difficiles à quantifier, nous préférons donc nous appuyer sur la notion de précurseurs d'évènements pour obtenir une définition plus précise, en considérant un évènement comme un pic d'activité et un précurseur comme un signal moins important ou moins intense, ayant une relation causale avec l'évènement.

1. Ce travail est soutenu par le programme "Investissements d'Avenir", projet ISITE-BFC (contrat ANR 15-IDEX-0003), <https://projet-cocktail.fr/>

2. <https://www.nap.edu/read/11061/chapter/6/#80>

Dans ce qui suit, nous présentons plusieurs études liées à nos travaux. Nous pouvons classer ces études en trois catégories : 1) Le « text mining » et le traitement du langage naturel (NLP); 2) L'apprentissage automatique (*Machine Learning ML*); et 3) les motifs ou patterns. Plusieurs approches de « text mining » et de NLP ont été proposées, dans lesquelles les documents Web sont étudiés à travers une analyse quantitative de mots-clés. Yoon (2012) a proposé deux indicateurs – le degré de visibilité basé sur la fréquence des mots-clés et le degré de diffusion basé sur la fréquence des documents – et a tenu compte de leurs taux d'augmentation dans le temps. Un mot clé ayant une faible visibilité et un faible niveau de diffusion est considéré comme un signal faible.

Ning et al. (2016) ont développé un algorithme d'apprentissage à instances multiples (*Multiple Instance Learning MIL*), basé sur des techniques d'apprentissage supervisé, afin de formuler le problème d'identification et de prévision des précurseurs. Le modèle consiste à attribuer une probabilité à des articles de presse traitant d'évènements à venir comme une manifestation. Une probabilité élevée est attribuée à un article si celui-ci est considéré comme un précurseur, contenant des informations sur les causes de cet évènement. Les auteurs ont montré, à travers trois jeux de données, que leur méthode était capable de prédire des manifestations. Une autre étude d'Ackley et al. (2020) a adopté des techniques d'apprentissage supervisé similaires dans le domaine de l'aviation, afin d'analyser et de suivre les paramètres critiques menant à des évènements de sécurité dans les phases d'approche et d'atterrissage des avions.

Certains chercheurs se sont intéressés aussi à l'identification de patterns spécifiques dans les réseaux, appelés motifs, qui pourraient être considérés comme des précurseurs d'évènements. Baiesi (2006) a présenté une méthode qui étudie les corrélations entre des graphes issus des tremblements de terre, en utilisant des outils sur la théorie des réseaux. Il a mesuré la distance entre les nœuds du réseau ainsi que leur coefficient de clustering. Après avoir appliqué des outils statistiques sur la topologie du réseau, il a constaté que des motifs simples tels que les triangles, constituent un type intéressant de précurseurs d'évènements majeurs car ils sont retrouvés dans les trois tremblements de terre étudiés. D'autres approches ont étudié l'identification et le rôle des motifs dans des évènements critiques tels que l'analyse de crimes (Davies et Marchione, 2015) et la détection d'attaques dans un réseau de télécommunication (Juszczyszyn et Kołaczek, 2011). Ces derniers travaux confortent l'hypothèse selon laquelle les graphlets, qui sont des motifs particuliers, peuvent être des précurseurs d'évènements.

3 Étude des graphlets comme précurseurs d'évènements

Les méthodes conventionnelles basées sur des techniques statistiques simples ne permettent pas d'identifier les précurseurs d'évènements dans les réseaux sociaux, elles sont plutôt utiles pour identifier des évènements tels que la famille des algorithmes ARIMA, EDM, HDC (Ray et al., 2018). Notre hypothèse est alors que la topologie du réseau joue un rôle important dans la propagation de l'information et nous choisissons d'explorer une approche basée sur la notion de graphlets. Les graphlets sont des types de motifs particuliers dans un réseau, ayant des tailles petites allant de 2 jusqu'à 5 nœuds. Grâce à leurs tailles et à leurs formes prédéfinies (30 types de graphlets), ils sont faciles à interpréter par des experts du domaine. Nous développons l'hypothèse que certains graphlets sont des précurseurs d'évènement potentiels tout comme les cliques le sont pour les communautés. Une fois les précurseurs potentiels révélés, il faut encore

valider le fait qu'il s'agit de signaux faibles, déterminer leur lien avec l'évènement étudié pour ensuite comprendre leur rôle.

Dans cette section, nous examinons les questions suivantes : Les graphlets peuvent-ils être utilisés pour identifier des précurseurs d'évènements ? Les mêmes types de graphlets peuvent-ils être observés avant les évènements ? Avant de rentrer dans les détails, nous présentons la notion de graphlets et nous détaillons les étapes de notre méthode.

3.1 Les graphlets, en quelques mots

Les graphlets ont été introduits pour la première fois par Pržulj et al. (2004). Un graphlet est un sous-graphe (2 à 5 nœuds) non isomorphe induit connecté choisi parmi les nœuds d'un grand graphe. Il existe 30 graphlets de G_0 à G_{29} jusqu'à 5 nœuds : le graphlet G_0 $\bullet \rightarrow$ de taille 2, deux graphlets de taille 3 qui sont G_1 $\bullet \rightarrow \bullet$ et G_2 \triangle , 6 graphlets de taille 4 et 21 graphlets de taille 5.

Les orbites, ou positions, représentent les classes d'équivalence des graphlets (Pržulj, 2007). Les nœuds appartenant à une même orbite sont interchangeables. Le degré des nœuds est généralisé ici à un vecteur de 73 coordonnées représentant les orbites des 30 types de graphlets.

Par exemple, le graphlet G_4 qui est en forme d'étoile $\begin{matrix} \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \end{matrix}$, possède deux positions, la première est centrale (orbite 7) occupée par un sommet, et la deuxième est périphérique (orbite 6) et partagée par les trois autres sommets.

Il existe plusieurs algorithmes qui énumèrent les graphlets et les orbites d'un graphe tels que RAGE, FANMOD, GraphCrunch et Orca (Ribeiro et al., 2019). Certains d'entre eux sont limités aux graphlets à 4 nœuds (par exemple RAGE), mais sont très efficaces pour les grands graphes. Nous nous appuyons sur l'algorithme Orca proposé par Hočevar et Demšar (2014), car il énumère les graphlets jusqu'à 5 nœuds tout en étant performant sur de grands graphes.

3.2 Méthodologie

L'identification de précurseurs d'évènements demande une série temporelle dans laquelle un évènement est vu comme un pic d'activité, et un précurseur est un signal d'intensité plus faible ayant une relation causale avec l'évènement.

Nous présentons dans la suite notre méthodologie constituée de six étapes, résumée en figure 1.

0. La première étape consiste à construire une série temporelle à partir de données issues des réseaux sociaux. Une méthode³ pour supprimer la saisonnalité de la série temporelle est ensuite appliquée. Une fois les données brutes collectées pour un intervalle de temps étudié, par exemple des tweets au format JSON, certaines interactions sont sélectionnées (comme retweet, citation, mention). Des tuples à 4 composantes représentant les interactions entre les entités à une date donnée sont générés, par exemple (`utilisateur1`, `utilisateur2`, `retweet`, `124354432`).

Une série temporelle X est construite à partir du nombre d'interactions sélectionnées entre toutes les paires de nœuds : $X = [x_1, x_2, \dots, x_n]$. Lorsqu'un pic d'activité est identifié, la série temporelle est divisée avant et pendant l'évènement en clichés S^t , par

3. <https://otexts.com/fpp2/x11.html>

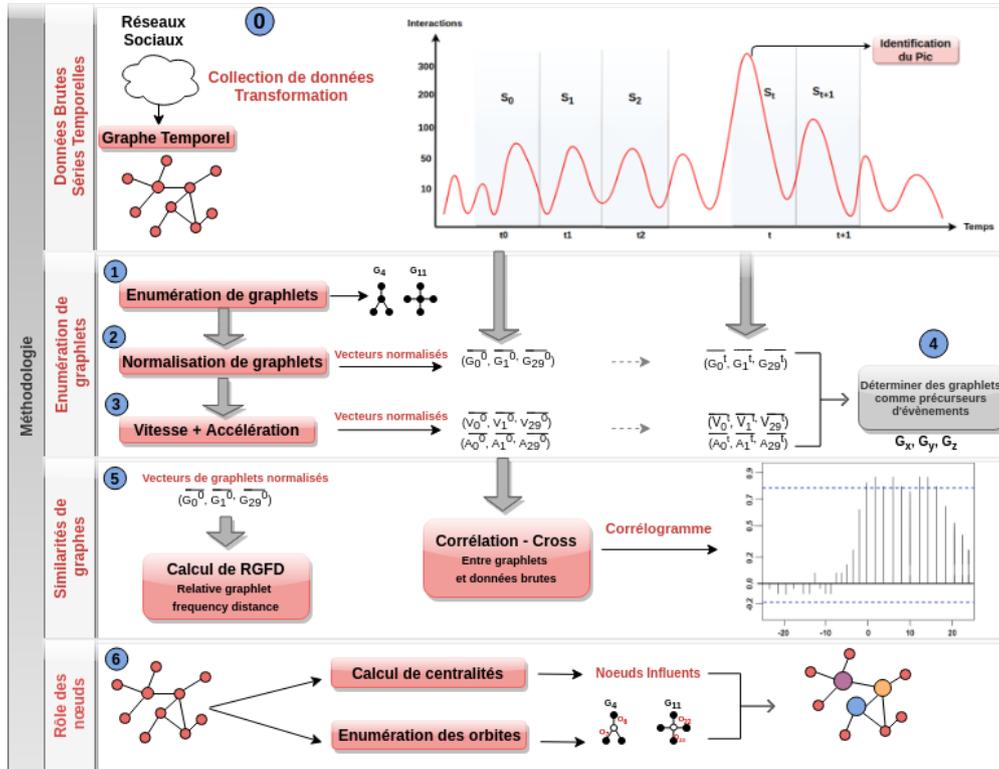


FIG. 1: Aperçu de la méthodologie

pas d'un jour, de 12 heures, de 6 heures, d'une heure, de façon à obtenir des sous-séries temporelles de la série d'origine : $S^t = [x_t, \dots, x_{t+1}]$ avec $S^t \subset X$.

1. Pour déterminer une signature topologique avant et pendant l'évènement, les graphlets de 2 à 5 nœuds sont énumérés pour chaque cliché⁴. Le temps de calcul prend au maximum quelques secondes. Chaque cliché S^t est représenté par un vecteur numérique $(G_0^t, G_1^t, \dots, G_{29}^t)$ où G_x^t est le nombre de graphlets de type x apparaissant dans le cliché S^t , pour x allant de 0 à 29.
2. Une procédure de normalisation est ensuite appliquée sur ces vecteurs afin de redimensionner leurs valeurs à une grandeur particulière, pour pouvoir appliquer des mesures et des calculs supplémentaires. Cette étape est d'une importance majeure car elle ne doit pas masquer les signaux faibles, mais plutôt les rendre comparables à d'autres. La procédure choisie est celle proposée par Goldin et Kanellakis (1995) dans laquelle ils étudient la similarité entre deux requêtes portant sur une base de données temporelles. Une requête renvoie une séquence X de nombres réels (x_1, \dots, x_n) . Deux réels a, b définissent une transformation $T_{a,b}$ sur X en mettant en relation chaque x_i avec $a * x_i + b$.

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

Détection de précurseurs d'évènements dans les réseaux sociaux

\bar{X} représente la forme normale de X , calculée par $\bar{X} = T_{\sigma, \mu}^{-1}(X) = T_{\frac{1}{\sigma}, -\frac{\mu}{\sigma}}(X)$, où $\mu(\bar{X}) = 0$ et $\sigma(\bar{X}) = 1$, μ étant la moyenne et σ la déviation standard.

En appliquant cette procédure de normalisation pour chacun des clichés S^t , chaque composante de son vecteur G_x^t est normalisée par :

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

où, pour tous les clichés, $\mu(G_x)$ est la moyenne et $\sigma(G_x)$ la déviation standard pour chaque type de graphlet x .

3. À partir des valeurs normalisées obtenues, leur vitesse $\overline{V_x^t}$ et leur accélération $\overline{A_x^t}$ sont calculées par $\overline{V_x^t} = \overline{G_x^{t+1}} - \overline{G_x^t} \forall x \in \{0, \dots, 29\}$ et $\overline{A_x^t} = \overline{V_x^{t+1}} - \overline{V_x^t} \forall x \in \{0, \dots, 29\}$
4. Les résultats obtenus dans les étapes 2 et 3 sont observés, afin de détecter des variations significatives dans leurs valeurs. Nous choisissons les k graphlets ayant les plus grandes valeurs de vitesse et d'accélération comme précurseurs potentiels d'évènements.
5. Afin de confirmer que les graphlets sont des précurseurs d'évènements, des méthodes d'analyse complémentaires sont appliquées, telles que :

Cross-Correlation sert à valider les propriétés intrinsèques de la méthodologie, sans pour autant établir une preuve de causalité. Il s'agit d'une mesure linéaire de similarités⁵ entre deux séries temporelles x et y qui permet d'évaluer la relation causale entre deux séries au cours du temps (Ripley et Venables, 2002). Un décalage h est associé à cette mesure, sachant que si $h < 0$ alors x prédit y , et si $h > 0$ alors y prédit x . Nous utilisons cette corrélation pour identifier quels types de graphlets influencent ou prédisent la série temporelle d'origine calculée à l'étape 0.

Relative Graphlet Frequency Distance (RGFD) est une mesure de similarité de la structure locale entre deux graphes proposée par Pržulj et al. (2004). Dans notre cas, une distance élevée indique une différence importante entre deux clichés S^t et S^{t+1} , confirmant la survenue d'un évènement remarquable. Soit $T(G^t) = \sum_{x=1}^{29} (G_x^t)$ le nombre total de graphlets apparaissant dans le cliché S^t , la fréquence relative des graphlets $F_x(G^t)$ dans un cliché S^t pour un type de graphlet x est calculé par $F_x(G^t) = -\log((G_x^t)/T(G^t))$ ⁶. La distance de la fréquence relative entre deux clichés consécutifs S^t et S^{t+1} est ensuite définie par :

$$D(S^t, S^{t+1}) = \sum_{x=1}^{29} |F_x(G^t) - F_x(G^{t+1})|$$

6. Une étape complémentaire est réalisée pour contextualiser les résultats par l'étude du rôle des nœuds influents dans la survenue de l'évènement. Tout d'abord, les nœuds influents sont identifiés dans le graphe d'origine, à l'aide d'algorithmes de centralité comme Page Rank, k-core ou le degré de centralité. Parallèlement, une étude plus fine est effectuée pour analyser les positions (orbites) de tous les nœuds dans les graphlets.

5. implantée dans le package *tseries* de R : <https://www.rdocumentation.org/packages/tseries/versions/0.1-2/topics/ccf>

6. Le logarithme est utilisé pour lisser les proportions de graphlets.

Nous comptons, tous clichés confondus, le nombre de fois où chaque nœud apparaît dans les 73 orbites des 30 types de graphlets étudiés. Les résultats sont analysés afin de déterminer si les nœuds influents participent fortement dans les graphlets précurseurs d'évènements et à quelle position. Cette analyse donne une idée du rôle des nœuds influents dans l'émergence de l'évènement. Par exemple, si un nœud u se trouve majoritairement dans la position centrale du graphlet G_4 il peut être considéré comme diffuseur d'information, s'il se trouve dans la position centrale du graphlet G_{27} il peut être à l'origine d'une communauté émergente.

4 Description des données

Dans cette section, nous décrivons les ensembles de données utilisés pour nos expériences. Le premier ensemble de données étudie un évènement réel et les deux autres ensembles⁷ sont des réseaux sociaux benchmark utilisés pour confirmer nos résultats d'analyse.

Le **réseau Twitter (Incendie Lubrizol)** comporte des tweets publiés à la suite de l'incendie à l'usine Lubrizol à Rouen, France. Nous choisissons de travailler par pas de une heure afin de découvrir des motifs cachés ou non remarquables dans le graphe, qui pourraient être invisibles si de grands intervalles de temps étaient utilisés. L'évènement sélectionné est la visite du Président de la République à Rouen le 30 octobre 2019 à 18 heures. Le corpus est constitué des tweets publiés entre le 28 Octobre 2019 minuit et le 30 Octobre 2019 minuit, il contient 18 914 tweets, 2 028 de ces tweets sont des tweets originaux et 2 724 incluent des mentions qui est le type d'interaction choisi.

Le **réseau MathOverflow** contient des interactions temporelles extraites du site Stack-Exchange "Math-Overflow", consistant en trois types d'interactions (répondre, commenter une question, commenter une réponse). L'ensemble des données utilisé est extrait de l'échantillon original et se compose de 1 400 relations du 27 octobre 2010 au 30 octobre 2010.

Le **réseau Facebook**, représente un ensemble de publications sur le mur d'autres utilisateurs sur Facebook. L'échantillon de données a été collecté entre octobre 2004 et janvier 2009, pour l'expérience nous réduisons l'ensemble entre le 5 janvier 2009 et le 7 janvier 2009, soit 8 790 interactions entre utilisateurs.

5 Expérimentations et discussion

Dans nos expériences, nous nous intéressons particulièrement à la validation de la méthodologie.

5.1 Expérience sur les données du réseau Twitter - Incendie Lubrizol

Dans cet ensemble de données, nous choisissons comme période d'étude les deux jours (28 et 29 octobre) précédant la visite du Président Macron à Rouen, ainsi que le jour de sa visite (30 octobre) et de travailler sur des clichés d'une heure. À l'étape 4 nous constatons une augmentation du nombre de certains graphlets le 30 octobre à partir de 16 heures (la visite

7. <https://snap.stanford.edu/data/\#socnets>

Détection de précurseurs d'évènements dans les réseaux sociaux

du Président était inattendue et a eu lieu vers 18 heures), comme G_2 , G_5 , G_8 , G_{27} , G_{28} . À partir de 18 heures le nombre d'autres types de graphlets commence à augmenter comme G_{11} , G_{14}  et G_{22} . Après 21 heures les nombres commencent à diminuer jusqu'à ce qu'ils deviennent négatifs à la fin de la journée. Le calcul de la vitesse et de l'accélération des graphlets met également en évidence la prédominance des types de graphlets ci-dessus, puisque les valeurs correspondantes sont supérieures aux valeurs obtenues pour les autres types de graphlets, pour les clichés avant l'évènement. Le tableau 1 compare la variation du nombre, de la vitesse et de l'accélération des graphlets entre quelques types de graphlets significatifs, qui ont évolué à 16 heures avant le pic, et d'autres graphlets qui n'ont pas montré de variations remarquables pour les mêmes clichés. Les valeurs les plus élevées sont surlignées en vert, bleu et rouge. Nous constatons également que les nombres des graphlets G_2 et G_5 augmentent considérablement par rapport aux autres types de graphlets, le 28 octobre entre 9 h et 13 h, ce qui indique un changement remarquable dans la structure du graphe.

Graphlet	$S^t : 30/10 15h$			$S^{t+1} : 30/10 16h$			$S^{t+2} : 30/10 17h$		
	G_x^t	V_x^t	A_x^t	G_x^t	V_x^t	A_x^t	G_x^t	V_x^t	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
G11 	-0,1907	0,0005	0,0023	-0,1347	0,0066	0,0554	7,6527	7,3927	7,3861
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

TAB. 1: Comparaison du nombre normalisé, de la vitesse et de l'accélération des graphlets, trois heures avant l'évènement (Seuls quelques graphlets sont présentés).

Le tableau 2 représente le résultat du calcul de la *relative graphlet frequency distance* (RGFD) pour certains clichés de l'après-midi du 30 octobre, jour de l'évènement. Nous notons une croissance significative de la distance à partir de 15 heures, pour atteindre la valeur maximale à 18 heures (heure de l'évènement) et ensuite une décroissance, montrant que la topologie du réseau a évolué de manière significative.

Clichés du 30/10	14h	15h	16h	17h	18h	19h	20h
$D(S^t, S^{t+1})$	13,9398	45,5913	80,4377	101,4438	119,9910	117,3054	99,8948

TAB. 2: RGFD calculée pour des clichés du 30 octobre

Les résultats obtenus à partir du calcul de la RGFD vont dans le sens de la capacité d'une telle mesure à représenter l'évolution de l'occurrence des graphlets en utilisant une seule valeur.

Après avoir appliqué la méthode de cross-corrélation, nous trouvons majoritairement les mêmes types de graphlets cités précédemment, positivement corrélés avec la série temporelle brute des mentions, c'est-à-dire que ces graphlets influent la relation mention. Par exemple, G_6 est corrélé avec la série des mentions par un décalage d'une heure, avec une valeur de 0,81. Ainsi que pour G_8 , G_{11} , G_{14} et G_{27} , les corrélations ont un décalage positif d'une heure avec des valeurs comprises entre 0,75 et 0,85. Pour des types de graphlets n'ayant pas été identifiés comme précurseurs, nous ne trouvons pas de corrélations avec la série temporelle brute, et le décalage est dans ce cas négatif d'une ou de deux heures, comme pour G_3 et G_{12} .

Nous avons répété la méthodologie, en découpant la période de l'étude en clichés de 30 minutes, puis de 15 minutes, pour rechercher tous les précurseurs possibles et supplémentaires dans un intervalle de temps plus précis. Après cette expérimentation, nous trouvons de nouveau les graphlets G_2 , G_5 et G_{27} émergeant à 16 heures et à 16h30, et des graphlets tels que G_{11} et G_{22} apparaissant à partir de 17h30.

Ensuite, nous contextualisons les résultats en passant à une expérience de granularité plus fine, en comptant le nombre d'orbites pour chaque nœud du graphe d'origine, pour chaque cliché d'une heure. En parallèle, les algorithmes de centralité font sortir les nœuds les plus centraux / influents du graphe. Nous analysons ensuite la position de ces nœuds dans chaque type de graphlet. Nous trouvons des utilisateurs comme `manon_leterq` un journaliste, `76actu` le site des informations local et `OTT_44380` qui ont le plus grand nombre d'orbites sur les périphéries (O_{22} et O_{32}) des graphlets G_{11} et G_{14} , en plus des utilisateurs `BFMTV` et `paris_normandie` avec un nombre remarquable d'orbites dans le même intervalle de temps.

Par exemple un nœud avec une position centrale dans le graphlet G_{11}  est directement connecté à 4 autres nœuds qui sont déconnectés les uns des autres. La position périphérique de ces nœuds trouvés et leur rôle dans la vie réelle, montrent qu'ils sont de bons candidats pour diffuser de l'information, alors qu'une position centrale indique que le nœud est un initiateur de l'évènement. Ces utilisateurs se retrouvent ainsi parmi les premiers utilisateurs identifiés par les algorithmes de centralités. Le tableau 3 présente une comparaison entre les nombres d'orbites pour les nœuds les plus influents obtenus à partir du Page Rank. Les nœuds les plus influents apparaissent principalement dans des graphlets identifiés comme précurseurs d'évènements.

Utilisateur	Page Rank	G11		G14		
		O22 	O23 	O31 	O32 	O33 
<code>paris_normandie</code>	1	112 663	74	3 445	15 189	24
<code>BFMTV</code>	2	99 173	80	3 119	10 617	30
<code>OTT_44380</code>	3	96 791	280	2 661	13 821	129
<code>76actu</code>	4	84 413	165	1 364	25 710	192
<code>Manon_Leterq</code>	5	69 101	470	1 198	36 128	135

TAB. 3: Page Rank et nombre de fois où un nœud (utilisateur) apparaît dans un orbite, pour tous les clichés (Extrait).

5.2 Expérience sur les réseaux Benchmark

La même méthodologie est appliquée sur les deux réseaux benchmark. Nous travaillons également par clichés d'une heure. Les résultats observés dans MathOverFlow montrent des variations significatives de certains nombres de graphlets le 29 octobre à partir de 22 heures.

Nous trouvons les graphlets G_3 , G_4 , G_9 et G_{10} augmentant d'abord, puis G_2 à 23 heures, heure du pic de l'activité. Ensuite, les nombres commencent à diminuer. Une autre montée significative intervient le 28 octobre à 1 heure du matin, pour les graphlets

G_6 , G_7 , G_{12} , G_{13} , G_{17} et G_{19} . Cette augmentation met en évidence un changement important dans la structure du graphe à ce moment. Après application de la méthode cross-corrélation, nous trouvons une corrélation positive avec un décalage de 5 heures, et des valeurs moyennes entre $+(-)0.3$ et 0.5 maximum, pour les graphlets cités. D'autre part, nous comptons le nombre d'orbites pour tous les nœuds du graphe d'origine, et extrayons en parallèle les nœuds les plus influents à l'aide des algorithmes de centralités. Nous constatons que la plupart des nœuds étant parmi ceux classés en premier, se trouvent dans des positions centrales, les nœuds restants se trouvent dans des positions périphériques.

L'ensemble de données du réseau Facebook a aussi été étudié par clichés d'une heure, le pic d'activité dans cet ensemble correspond à un évènement survenu le 7 janvier 2009 à partir de 6 heures. La même méthodologie est appliquée à ce corpus, d'où la montée du nombre des graphlets G_3 et G_9 en même temps que le pic du 07 janvier à 6 heures du matin. Nous remarquons également une augmentation notable du nombre de graphlet G_2 durant la soirée avant le pic à 18 heures. Ces types de graphlets représentent des communautés grâce à leur forme triangulaire. De plus, l'énumération des orbites indique que les utilisateurs classés en premier avec les algorithmes de centralités, ont les valeurs les plus élevées en orbites, correspondant à des positions centrales plutôt qu'à des positions périphériques ou foliaires.

Le tableau 4 présente une comparaison du nombre des nœuds dans les trois réseaux, pour les orbites de certains graphlets que nous considérons comme précurseurs d'évènements. Dans les trois réseaux nous remarquons que la plupart des nœuds ont tendance à avoir des positions périphériques dans G_6 , G_9 et G_{11} , occupant respectivement les orbites O_{10} , O_{15} et O_{22} , plutôt que des positions centrales comme les orbites O_{11} , O_{17} et O_{23} .

Graphlets	G6			G9			G11	
								
Orbites	O9	O10	O11	O15	O16	O17	O22	O23
Twitter - Incendie Lubrizol	31 360	63 352	32 330	582 418	589 732	290 375	297 989 445	76 354 106
MathOverFlow	22	44	22	1050	1050	525	1712	428
Facebook	2	4	2	74	74	37	3 352	838

TAB. 4: Résultats de l'énumération des orbites dans les trois graphes, pour les graphlets G_6 , G_9 et G_{11} .

Nous avons effectué des analyses quantitatives et qualitatives à l'aide de plusieurs algorithmes d'énumération et de similarité de graphes. Les résultats expérimentaux obtenus permettent de confirmer notre hypothèse qui considère que les graphlets sont des précurseurs d'évènements et qu'ils peuvent être vus comme des signaux faibles.⁸

8. Les données et les programmes des expérimentations sont disponibles (<https://github.com/hibaaboujamra/EventPrecursorsGraphlets>).

6 Conclusion et perspectives

Plusieurs études ont analysé la propagation du signal dans de grands réseaux complexes, mais la plupart de ces études ne prennent pas en compte la structure locale du réseau. Nous avons étudié dans cet article le rôle de la structure locale dans la détection des signaux faibles dans les données des réseaux sociaux, en utilisant plusieurs ensembles de données. Nous avons posé comme postulat que les graphlets sont des précurseurs d'évènements, et nous utilisons des méthodes algorithmiques diverses pour évaluer et confirmer cette hypothèse. Les résultats ont mis en évidence certaines singularités avant les évènements étudiés, et des corrélations entre plusieurs types de graphlets et la série temporelle brute étudiée. Nous avons analysé également le rôle des nœuds dans la production de tels évènements.

Du point de vue expérimental, nous avons pu détecter des motifs spécifiques avant l'apparition d'évènements à partir des séries temporelles étudiées. Nous avons remarqué une présence significative de différents types de graphlets tels que G_2 , G_6 , G_9 , G_{10} et G_{11} , dans les trois réseaux étudiés. Nous avons constaté également à partir de ces expériences, et sur une période assez longue avant l'évènement, que le nombre de graphlets ne variait pas beaucoup. Une corrélation est aussi trouvée entre les précurseurs d'évènements détectés et l'évènement lui-même, et nous avons remarqué que certains graphlets dans les données de Lubrizol permettent une anticipation d'une heure. Nous nous attachons actuellement à établir une preuve plus formelle de la causalité en utilisant par exemple la causalité de Granger.

Dans les travaux futurs, nous souhaitons appliquer notre méthodologie sur des graphes orientés, à partir des travaux d'Aparício et al. (2019), dans lesquels ils mesurent la dominance directe et indirecte des nœuds. Nous souhaitons également tester la méthode dans un processus itératif consistant à éliminer les nœuds en continu des graphes, afin de diminuer la dominance de certains graphlets. Cela permettra d'obtenir une décomposition hiérarchique en graphlets, et de découvrir l'ensemble des graphlets qui apparaissent toujours comme précurseurs d'évènements quelque soit le réseau ou l'évènement.

Références

- Ackley, J. L., T. G. Puranik, et D. Mavris (2020). A supervised learning approach for safety event precursor identification in commercial aviation. In *AIAA Aviation Forum*, pp. 2880.
- Ansoff, H. I. (1975). Managing strategic surprise by response to weak signals. *California management review* 18(2), 21–33.
- Aparício, D., P. Ribeiro, F. Silva, et J. Silva (2019). Finding dominant nodes using graphlets. In *International Conference on Complex Networks and Their Applications*, pp. 77–89. Springer.
- Baiesi, M. (2006). Scaling and precursor motifs in earthquake networks. *Physica A : statistical mechanics and its applications* 360(2), 534–542.
- Davies, T. et E. Marchione (2015). Event networks and the identification of crime pattern motifs. *PloS one* 10(11), e0143638.
- Goldin, D. Q. et P. C. Kanellakis (1995). On similarity queries for time-series data : constraint specification and implementation. In *International Conference on Principles and Practice of Constraint Programming*, pp. 137–153. Springer.

- Harrysson, M., E. Métayer, et H. Sarrazin (2014). The strength of weak signals. *McKinsey Quarterly* 1, 14–17.
- Hočevár, T. et J. Demšar (2014). A combinatorial approach to graphlet counting. *Bioinformatics* 30(4), 559–565.
- Juszczyszyn, K. et G. Kołaczek (2011). Motif-based attack detection in network communication graphs. In *IFIP International Conference on Communications and Multimedia Security*, pp. 206–213. Springer.
- Lesca, H. et S. Blanco (2002). Contribution à la capacité d'anticipation des entreprises par la sensibilisation aux signaux faibles. *CIFEPME, HEC-Montréal-Québec*.
- Leskovec, J. et A. Krevl (2014). SNAP Datasets : Stanford large network dataset collection. <http://snap.stanford.edu/data>.
- Ning, Y., S. Muthiah, H. Rangwala, et N. Ramakrishnan (2016). Modeling precursors for event forecasting via nested multi-instance learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1095–1104.
- Pržulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics* 23(2), e177–e183.
- Pržulj, N., D. G. Corneil, et I. Jurisica (2004). Modeling interactome : scale-free or geometric ? *Bioinform.* 20(18), 3508–3515.
- Ray, S., D. S. McEvoy, S. Aaron, T.-T. Hickman, et A. Wright (2018). Using statistical anomaly detection models to find clinical decision support malfunctions. *Journal of the American Medical Informatics Association* 25(7), 862–871.
- Ribeiro, P., P. Paredes, M. E. Silva, D. Aparicio, et F. Silva (2019). A survey on subgraph counting : concepts, algorithms and applications to network motifs and graphlets. *arXiv preprint :1910.13011*.
- Ripley, B. D. et W. Venables (2002). *Modern applied statistics with S*. Springer.
- Schoemaker, P. et G. S. Day (2009). How to make sense of weak signals. *Leading Organizations : Perspectives for a New Era* 37.
- Yoon, J. (2012). Detecting weak signals for long-term business opportunities using text mining of web news. *Expert Systems with Applications* 39(16), 12543–12550.

Summary

The availability of social networks data seeks researchers' interest to develop algorithms and machine learning models to analyze user interactions and behaviors. These algorithms rely on network topology to represent structural changes and to detect remarkable precursors generally preceding major events. The approach presented in this article aims to study whether certain graphlets (specific patterns) can be considered as precursors of an event. We experiment the proposed method on three different sets of social networks data. We also study the role (position) of influential nodes in the graphlets, which have a central position in the global graph. After analyzing the results, we show that graphlets are considerable precursors of events.