ExCoDE: a Tool for Discovering and Visualizing Regions of Correlation in Dynamic Networks

Giulia Preti
University of Trento
Trento, Italy
gp@disi.unitn.eu

Polina Rozenshtein
Aalto University
Espoo, Finland
polina.rozenshtein@aalto.fi

Aristides Gionis
Aalto University
Espoo, Finland
aristides.gionis@aalto.fi

Yannis Velegrakis
Utrecht University
Utrecht, Netherlands
i.velegrakis@uu.nl

***Abstract*—Dynamic graphs are valuable means to represent the volatility of real-world networks. In such scenarios, dense subgraph mining is a widely studied task, as it can give insights about how the relationships change over time. However, there are cases in which these changes are correlated. For example, in a road network, a traffic accident affects also the traffic in the adjacent road segments. We consider the problem of detecting dense regions of correlation in dynamically evolving networks, and demonstrate EXCODE, a system that solves two variants of the problem, which are based on two different density measures. It enumerates all the subgraphs satisfying certain density and correlation constraints, but can also detect compact subsets of limited overlap. In this demonstration, the audience can try this tool with real-world datasets, hence visualizing, interacting, and exploring the dense correlated subgraphs discovered in the mining process. An interactive panel allows them to learn where the correlations are located in the network, how the regions of correlation are related to each other, and how they evolve over time.**

***Index Terms*—Dynamic Graphs; Dense Subgraphs; Correlated Subgraphs;**

I. INTRODUCTION

Graphs have nowadays attracted considerable attention due to their expressive power and flexibility. Many real life situations can be naturally modeled through dynamic networks, which are graphs that change over time. Analyzing only specific snapshots or aggregates of such networks leaves out a great deal of valuable information and insights that can be obtained by studying them over time. Thus, studying dynamic graphs by taking into consideration the whole evolution history is of paramount importance.

In the context of dynamic networks, many efforts have been devoted to dense subgraph mining, which aims at identifying portions of the network that are highly connected [1], [2], [3]. However, little work has focused on the detection of dense and correlated regions [4], which represent portions of the network characterized by similar structural or qualitative changes. This problem finds application in a wide range of scenarios such as fault diagnosis and root cause analysis [5], [6], as symptoms caused by the same root cause tend to have a similar behavior and be located close in the network.

Example. *The BGP protocol establishes how the routers forward the packets across the Internet. Managing the Internet network requires, among others, the detection of issues in the BGP routing topology and then the diagnosis of their causes. This analysis allows a faster recovery and may prevent these problems from happening again. The BGP routing topology can be modeled ad a dynamic graph where nodes are routers and edges are routing paths. The edges are dynamic because the corresponding paths can change due to reconfigurations or faults. A fault at some router can induce changes in other portions of the graph, as all the paths traversing the faulty router are affected and must be replaced to ensure the continuation of the routing operations. As a consequence, changes in the same periods of time involving a group of edges close in the graph are likely caused by the same cause. Therefore, by focusing the attention on a maximal dense group of temporally correlated routes, the network manager can isolate the root cause of their faults more easily. However, in each snapshot of the whole BGP graph there can be a significant number of elements experiencing a change that need to be analyzed by the manager, and in addition, not every change is associated with an anomalous event. Thus, there is a need for an automatic tool that can simplify the detection of the issues by finding the regions in the graph whose edges present a similar pattern of appearance, so that the analyst need to focus on a small number of network elements.*

In this demonstration, we showcase EXCODE, a general framework for finding dense correlated subgraphs in dynamic networks. This framework uses two different measures to compute the density of a group of edges that change over time, and a measure based on the Pearson correlation to compute their correlation. To deal with the problem of huge result sets peculiar to tasks that enumerate all the solutions satisfying given constraints, the system is also able to select and return a compact but yet informative subset of solutions. This subset contains maximal and highly diverse subgraphs that, all together, are representative of the whole answer set.

II. TECHNICAL BACKGROUND

The input is a *dynamic network*, which is a graph modeled as a sequence of static graphs called *snapshots*:

Definition 1: A *dynamic network* $D = (V, E)$ is a sequence of graphs $G_i = (V, E_i, \omega_i)$ with $i \in T$, where V is a set of vertices, $E_i \subseteq V \times V$ is a set of edges between vertices, $\omega_i : E_i \mapsto \mathbb{R}$ is an edge weight function, and T is a set of time instances. The union of the edges of the snapshots is denoted by E, i.e., $E = \cup_{i \in T} E_i$.

We consider networks where all the snapshots share the same set of nodes, and assume that $\omega_i(e) = 0$ if e does not appear in G_i. When the snapshots are unweighted, the edge weights ω_i take values in $\{0, 1\}$.

The goal of ExCoDE is to extract the dense correlated subgraphs from D with size smaller than a given threshold s_{max}. Given the network D, a *subgraph* H of D is a graph $H = (V_H, E_H)$, such that $V_H \subseteq V$ and $E_H \subseteq E$. The density of a subgraph H in a static graph is generally defined as the average degree of its nodes, i.e., $\rho(H) = 2|E_H|/|V_H|$. However, when the graph is dynamic, the edges of H may not always exist, and thus the average node degree of H changes over time. Therefore, ExCoDE uses two approaches to aggregate those degrees and obtain a single density score. The first approach, called *minimum density* and denoted as ρ_m^k, computes the density of H as the minimum density of any subgraph induced by H across the snapshots where at least k edges of H are present. The second approach, called *average density* and denoted as ρ_a^k, computes the average density. Let $G_i(H) = (V_H, E_H \cap E_i)$ denote the subgraph induced by H in the snapshot i, and T_H^k denote the subset of snapshots where at least k edges of H appear, i.e., $T_H^k = \{t \mid t \in T \text{ and } |E_t \cap E_H| \geq k\}$. Then,

1) $\rho_m^k(H) = \min_{i \in T_H^k} \rho(G_i(H))$

2) $\rho_a^k(H) = 1/\left|T_H^k\right| \sum_{i \in T_H^k} \rho(G_i(H))$

If T_H^k is empty then both $\rho_m^k(H)$ and $\rho_a^k(H)$ are defined to be 0. Given a density threshold δ, a subgraph H is *dense* if $\rho^k(H) \geq \delta$, where ρ^k collectively indicates ρ_m^k and ρ_a^k. For each snapshot $t \in T_H^k$, we say that H is *active* in t.

Note that the set T_H^k is used to account for situations where H is highly dense in some snapshots, but it does not appear in other snapshots. For these particular subgraphs, the minimum density over all the snapshots would be 0 even if there is just one snapshot in which H does not appear, and the average density will take low values, even if H has a large average degree in all the snapshot in which it appears. Using T_H^k, all the snapshots where H is absent or where only few of its edges are present are discarded, hence obtaining larger values of minimum and average density.

The *correlation of* H is computed in terms of the pairwise correlation of its edges. Intuitively, two edges are correlated if they present a similar pattern of appearance over the snapshots of the network. Let $\mathbf{t}(e) = \{t_1(e), \ldots, t_T(e)\}$ denote the sequence of weights of the edge e, i.e., $t_i(e) = \omega_i(e)$ for $i \in T$. Then, the correlation $c(e_1, e_2)$ between two edges e_1 and e_2 is the Pearson correlation between $t(e_1)$ and $t(e_2)$. The correlation of H is defined as the minimum pairwise correlation among its edges, i.e., $c_m(H) = \min_{e_i \neq e_j \in E_H} c(e_i, e_j)$, and given a correlation threshold σ, H is *correlated* if $c_m(H) \geq \sigma$.

Since a dense correlated subgraph may contain smaller dense correlated structures due to the nature of the density and correlation measures, we reduce the size of the answer set by focusing on the *maximal diverse* subgraphs, which are subgraphs that are not contained in other dense correlated subgraphs and that differ from one another. We calculate the similarity between two subgraphs G'=(V', E') and G''=(V'', E'') using the Jaccard similarity between their edge sets, i.e., $J(G', G'')$=$|E' \cap E''|/|E' \cup E''|$, and we require that the pairwise similarities are lower than a given similarity threshold ϵ.

III. System Overview

ExCoDE first identifies the maximal sets of correlated edges, and then extracts subsets of these edges that form a maximal dense subgraph according to one of the two density measure ρ_m^k or ρ_a^k. Given a dynamic network $D = (V, E)$, it creates a graph $\mathcal{G} = (E, \mathcal{E})$ where the nodes are the edges E of D, and the edges are the pairs $(e_1, e_2) \in E \times E$ with correlation $c(e_1, e_2) \geq \sigma$. With this construction, a *maximal clique* in $\mathcal{G}$ corresponds to a maximal set of correlated edges in D, because a set of nodes forms a clique if and only if the nodes are all connected, and in this case the corresponding edges in D have pairwise correlation greater than σ. Then, given the maximal groups of correlated edges $\mathcal{C}$, ExCoDE examines each connected component in $\mathcal{C}$ to identify those constituting maximal dense subgraphs in D, retaining only a subset of pairwise dissimilar subgraphs according to the similarity threshold ϵ.

Enumeration of the maximal correlated edges. The graph $\mathcal{G}$ is built by computing the correlation $c(e_1, e_2)$ between each pair of edges $e_1, e_2 \in E$ and retaining those pairs satisfying $c(e_1, e_2) \geq \sigma$. The maximal groups of correlated edges are enumerated using the GP algorithm [7] for maximal clique discovery.

Extraction of the dense subgraphs. Given the maximal groups of correlated edges $\mathcal{C}$, ExCoDE needs to extract the groups of edges that form a maximal dense subgraph, using either ρ_m^k or ρ_a^k as density function. Since the edges in a group are not necessarily connected in D, it first extracts all the connected components from each group. To allow a faster discovery of the maximal groups of dense edges, the connected components are sorted in descending order of their size and processed iteratively.

If no dense set larger or similar to the current candidate X has been discovered yet, and if the size of X does not exceed the threshold s_{max}, the density of X is computed applying either Equation **1)** or Equation **2)**. When the density is above the threshold δ, X is inserted in the result set $\mathcal{S}$. When the density is below the threshold δ, the set X is not dense; though some subset $X' \subseteq X$ may be dense. Therefore, ExCoDE uses an approach based on a 2-approximation algorithm for the densest subgraph problem [8] to extract the dense subsets in X. These subsets are inserted into an auxiliary set $\mathcal{P}$. When all the candidates have been examined, the maximal groups in $\mathcal{P}$ that are not similar to any edge set in $\mathcal{S}$, are finally added to $\mathcal{S}$.

We evaluated the efficiency and effectiveness of ExCoDE with an extensive set of experiments on both real and synthetic datasets of increasing size. Figure 1 shows the running time

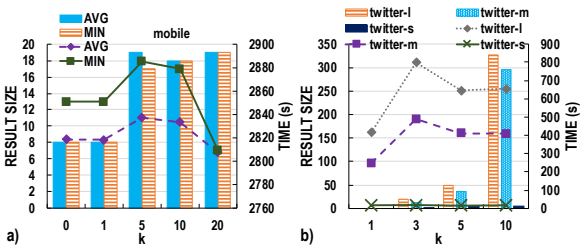

Fig. 1. Running time and number of subgraphs extracted from a mobile network (a) and three hashtag networks (b), varying the threshold k.

and the size of the answer when varying the edge-per-snapshot threshold k in a mobile communication network of size $(|V|, |E|, |T|) = (5K, 80K, 48)$ (a), and in three samples of a hashtag co-occurrence network of size (767, 2K, 2K), (1.2K, 7K, 2K), and (1.3K, 10K, 2K), respectively.

IV. Demonstration

Goals In this demonstration, the attendees will test the effectiveness of ExCoDE in mining maximal dense correlated subgraphs in dynamic networks, and observe and examine its outcomes. Therefore, the goal is two-fold: (i) to recognize the value that this kind of analysis tool can bring and the convenience of the results produced in real-world applications, and (ii) to understand how the system works, how it overcomes the computational challenges, and how its parameters affect the output.

The system gives the possibility of characterizing the subgraphs extracted from the network, so that they better match the expectations of the user. Then, an interactive panel displays the subgraphs discovered, allowing the users to visualize where the regions of correlation are located, how strong they are, and how they are related with each other. An additional panel enables the analysis of each region separately, and in particular, the users can investigate how each region behaves over time, how its density changes, and in which snapshots it is active.

Audience This demonstration is directed at any data practitioner who needs a tool to visualize and interact with large dynamic networks, and would like to understand how the network evolves, or more specifically, identify unexpected and significant events happening in the network. It is also interesting for any data engineer who needs an effective tool to detect correlated anomalies in the network under supervision, and any data researcher who is curious about the challenges behind understanding how a given dynamic network behaves over time, and how its changes are related to each other.

Scenario The demonstration starts with loading the dataset and, optionally, a mapping file containing the node labels (Figure 2 (a)). Among others, we consider a sequence of snapshots of the Internet network topology created with the routing tables used from August 29 to August 31, 2005. In these days, a catastrophic hurricane hit Florida and Louisiana, causing major issues also to the routing topology. Once the

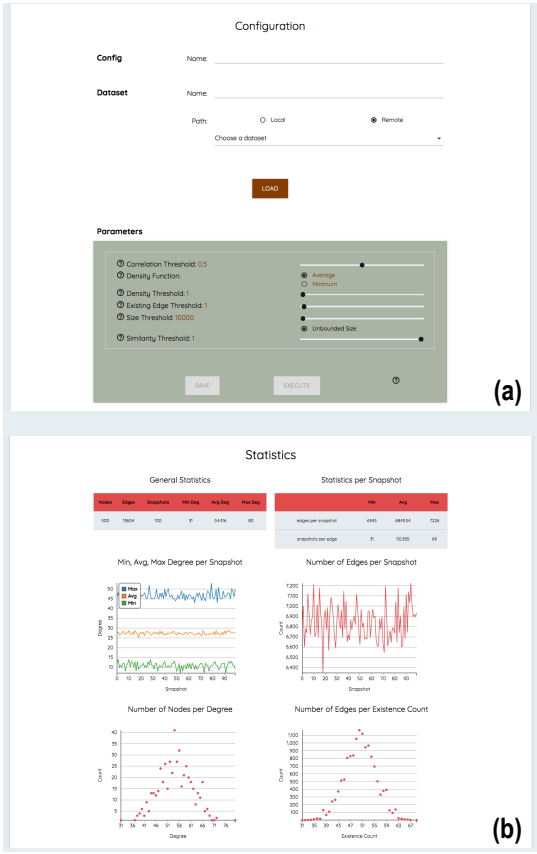

Fig. 2. Dataset selection (a) and dataset statistics (b).

dataset is loaded, a visualization of its main characteristics, such as min/avg/max degree, number of edges over time, degree distribution, and edge distribution, is presented to the audience (Figure 2 (b)). Then, the system guides the users towards configuring the parameters of the system according to their desires. These parameters involve desired levels of density, correlation, size, activity, and redundancy in the results. The system explains the role of each parameter, what values it can take, and how such values affect the output. This information, together with the insights provided by the charts displaying the graph characteristics, can help them to select appropriate values.

Once the parameters are configured, the mining algorithm is executed, and the results are given to the users. An interactive panel shows the graph with the dense groups of correlated edges highlighted using different colors, as illustrated in Figure 3(c). The edges in each group are characterized by a similar behavior over time, and are topologically close in the snapshots where the group is active. Denser subgraphs are colored in darker colors, nodes belonging to multiple subgraphs are indicated in black, and nodes that are not part of any dense subgraph are in white. The users can interact with the graph to better understand its structure. For example, by hovering over a node, they can see to which subgraphs the node belongs, and by clicking on it they can select that

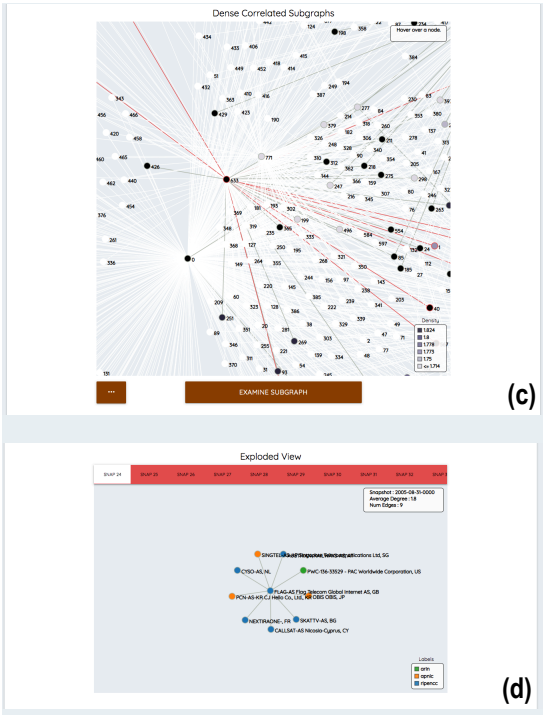

Fig. 3. Dense correlated subgraphs (c) and exploration of a subgraph (d).

(or those) subgraph. In addition, they can drag the nodes, and zoom in and out of the graph. Further information on the dense subgraphs is provided in a separate hidden panel, allowing them to understand where the regions of correlation are located in the network, as well as how they are related with each other. Finally, the users can explore and analyze a dense subgraph in isolation, by selecting it in the main panel and clicking the *examine subgraph* button. A separate panel (Figure 3 (d)) shows how the subgraph changes over time, how many edges appears in each snapshot where the subgraph is active, and the average degree. Different colors are used to highlight the different types of nodes, according to the mapping file loaded at the beginning of the demonstration. For example, in the case of the Internet network topology, the colors indicate the regional Internet address registries for the five geographical areas: Asia-pacific, North America, South America, Europe, and Africa. Thanks to this tool, a network analyst can see which countries were affected by the disaster and which issues were caused by the same root cause, sparing him the trouble of examining each node in the network.

V. Related work

This work is related to dense subgraph mining in dynamic networks [9], [3], and in particular, to the problem of enumerating all the dense structures satisfying given constraints [10], [11], [12], [13]. However, all these works retrieve either the single best solution or dense subgraphs whose edges are not temporally correlated. Additional measures of interestingness have been considered in fraud detection, with the goal of finding suspicious regions. These works measure the suspiciousness as the negative log likelihood according to a Poisson distribution [14], [15], the difference in density with the previous snapshots [16], or the sum of the anomaly scores of the nodes and edges [17]. Although they propose more complex measures, they do not detect groups of edges with similar behavior over time. A notion of correlation has been introduced by Guan et al. [18] and Yu et al. [19], which, however, assign a label to each node and retrieve those nodes that unusually deviate during some time interval. The most related approaches to this work [20], [4] find regions of correlated temporal change in dynamic graphs by expressing the temporal similarity between two edges as the Euclidean similarity between their time series, and the spatial similarity as the shortest path distance. They iterate over the snapshots of the graph using a window of fixed size, and for each window, they hard partition all the edges using first the temporal distance and then the spatial distance. In contrast, we enumerate only the groups of edges with large density and high pairwise edge correlation.

References

[1] A. Epasto, S. Lattanzi, and M. Sozio, "Efficient densest subgraph computation in evolving graphs," in *WWW*, 2015.

[2] Y. Yang, D. Yan, H. Wu, J. Cheng, S. Zhou, and J. Lui, "Diversified temporal subgraph pattern mining," in *SIGKDD*, 2016.

[3] P. Rozenshtein, N. Tatti, and A. Gionis, "Finding dynamic dense subgraphs," *TKDD*, vol. 11, no. 3, 2017.

[4] J. Chan, J. Bailey, C. Leckie, and M. Houle, "ciforager: Incrementally discovering regions of correlated change in evolving graphs," *TKDD*, vol. 6, no. 3, 2012.

[5] H. Kashima, T. Tsumura, T. Idé, T. Nogayama, R. Hirade, H. Etoh, and T. Fukuda, "Network-based problem detection for distributed systems," in *ICDE*, 2005.

[6] S. Kandula, D. Katabi, and J.-P. Vasseur, "Shrink: A tool for failure diagnosis in ip networks," in *SIGCOMM*, 2005.

[7] Z. Wang, Q. Chen, B. Hou, B. Suo, Z. Li, W. Pan, and Z. G. Ives, "Parallelizing maximal clique and k-plex enumeration over graph data," *J Parallel Distrib Comput.*, vol. 106, 2017.

[8] M. Charikar, "Greedy approximation algorithms for finding dense components in a graph," in *APPROX*, 2000.

[9] P. Bogdanov, M. Mongiovì, and A. K. Singh, "Mining heavy subgraphs in time-evolving networks," in *ICDM*, 2011.

[10] J. Pei, D. Jiang, and A. Zhang, "On mining cross-graph quasi-cliques," in *SIGKDD*, 2005.

[11] M.-S. Kim and J. Han, "A particle-and-density based evolutionary clustering method for dynamic networks," *VLDB*, vol. 2, no. 1, 2009.

[12] G. Qin, L. Gao, and J. Yang, "Significant substructure discovery in dynamic networks," *Current Bioinformatics*, vol. 8, no. 1, 2013.

[13] N. Shah, D. Koutra, T. Zou, B. Gallagher, and C. Faloutsos, "Timecrunch: Interpretable dynamic graph summarization," in *SIGKDD*, 2015.

[14] M. Jiang, A. Beutel, P. Cui, B. Hooi, S. Yang, and C. Faloutsos, "A general suspiciousness metric for dense blocks in multimodal data," in *ICDM*, 2015.

[15] K. Shin, B. Hooi, and C. Faloutsos, "M-zoom: Fast dense-block detection in tensors with quality guarantees," in *PKDD*, 2016.

[16] D. Eswaran, C. Faloutsos, S. Guha, and N. Mishra, "Spotlight: Detecting anomalies in streaming graphs," in *SIGKDD*, 2018.

[17] M. Mongiovi, P. Bogdanov, R. Ranca, E. E. Papalexakis, C. Faloutsos, and A. K. Singh, "Netspot: Spotting significant anomalous regions on dynamic networks," in *ICDM*, 2013.

[18] Z. Guan, X. Yan, and L. M. Kaplan, "Measuring two-event structural correlations on graphs," *PVLDB*, vol. 5, no. 11, 2012.

[19] W. Yu, C. C. Aggarwal, S. Ma, and H. Wang, "On anomalous hotspot discovery in graph streams," in *ICDM*, 2013.

[20] J. Chan, J. Bailey, and C. Leckie, "Discovering correlated spatio-temporal changes in evolving graphs," *KAIS*, vol. 16, no. 1, 2008.