Survey on Methods for Real-world contagion prediction based on Motif

Subin Kim

School of Computing, Korea Advanced Institute of Science and Technology
21supersoo@kaist.ac.kr

Abstract. Network is a useful tool for modeling various subjects, such as biology, physics, healthcare, etc. Finding the hidden patterns between data using network structures can give useful information. motifs are isomorphic sub-graphs that appear repeatedly throughout different parts of the network, involving different nodes. Using network motif as a tool for apprehending such patterns has been practiced. Here is the overview of using motif in analyzing the patterns of contagious diseases in networks and the plausibility of future research.

Keywords: Influence maximization · Network epidemics · network motif

1 Introduction

Networks are typically used to model systems that connect different types of entities. Classifying nodes and predicting edges are two essential tasks that many complex systems have tried to perform. Meanwhile, high-order structures as well as repetitive , smaller subgraph patterns are essential in the network[1]. This repeated subgraph pattern represents a motif. The concept of motifs was first proposed in the gene regulatory network. Then, a significant number of studies begin to focus on the computational theory of the network motif. To elaborate, network motifs are considered as network components of complex networks, which are described as a much more frequent connection pattern than expected[2]. Patterns between a few nodes in a network can be modeled as network motifs, which allow researchers to discover interesting properties of large-scale networks using motif discovery algorithms. They can be applied to predicting the progress of epidemic networks through grasping the structure of networks.

Mathematical modeling of epidemic transmission was progressing towards more powerful and flexible methods. The initial model focused on understanding the relationship between epidemiological parameters, such as the transmission rate, duration of infection, etc. Also, it focused on using simple measures of epidemic effectiveness such as the underlying reproductive number of diseases or the epidemic size on final stages. Questions such as how much vaccination should be done can be thrown in terms of measuring the the proportion of the population that was effected, or how much vaccine should be supplied in terms of overall

reduction in the spreading of disease across the entire network. There are also perceptions that some nodes are more important than others for transmission purposes, and therefore if one can control the key nodes, one can maximize the influence in the network. This perspective have been gaining support especially since the adoption of the original envisioned method in the context of social network analysis.

This survey tries to lighten the studies related with modeling epidemic spreading through motif. In **section 2**, the term of motif is explained, and in **section 3**, previous researches related with prediction for Real-world contagion based on motif are introduced. **section 4** contains descriptions of several research questions that can be derived from previous works introduced in **section 3**.

2 Terms of Motif and its property

Since motifs can characterize the dynamics and functional behaviors of networks, it can be applied to distinguish different types of networks based on motif statistics within the network the motif belongs to. Other researchers have found that there exist some particular patterns that can be modeled by certain network substructures in real-world networks. Such patterns have practical meanings, including social relationships, protein complex, information infrastructure, etc. These patterns are reflected by certain kinds of network substructures and appear much more often in real-world networks. In order to distinguish whether such kind of patterns are motifs or not, we can generate corresponding random networks and compare them with the original graph. Previously, graphlets are used to indicate structural patterns of networks, which encode diverse functional and relevant roles of different nodes. Statistically over-represented graphlets are defined as motif.

Definition 1. *Network motifs are patterns of interconnections that occur with significantly higher frequencies in complex networks than those in random networks.*

According to the definition 1, a subgraph is thought as a motif when it is represented - in a certain network - over a threshold with more than a number of times campared to randomly generated graph. A motif having k nodes is called k-motif, and the smallest motifs in directed networks is made up of two nodes. On the other hand, motif with three nodes is the most frequent structure in undirected networks. When generating random networks, it is essential to make the incoming and outgoing node degrees similar to those in the real networks. Also, it should be guaranteed that the frequencies of $(k-1)$-motifs of the original network and that of the randomized network are the same when searching the k-motifs. This is because the larger motif with k nodes can be decomposed into at least one motif with $(k-1)$ nodes. It ensures that the high significance of a target motif does not simply generate from its sub-motifs. The two criteria stated above maintain the intrinsic global and local properties of the original and randomized networks. We aim to ensure that the appearance of motifs in networks is not

determined by the properties of general networks including random graphs, but specific to the particular networks.

Definition 1 is one of the most widespread network motif definitions, which is also regarded as the standard definition of network motifs. However, this definition is proposed based on the topologies of networks and neglects the individual properties of different motifs. Therefore, because there are different types of networks, such as directed and undirected networks, we should try to seek different definitions under different scenarios.

Besides over-represented substructures, under-represented substructures in the networks are also meaningful for describing the network properties. An under-represented definition of network motifs, anti-motif[3] is proposed as the following:

Definition 2. *Sub-networks that are significantly underrepresented in the original networks are anti-motifs.*

Considering the sub-networks that induced from the original networks, the notion of induced motif has been proposed. An induced motif is a sub-network that is composed of all edges between its nodes in the original networks. Moreover, a strongly connected motif is a sub-network, in which there is a path from any node to any other node[4].

In gene regulatory networks, motifs are defined as small, repeated, and evolutionary conserved sub-networks. To be more specific, the transcription network illustrates the motifs with specific functionality in determining gene expressions, i.e., generating the temporal expression programs and controlling the responses to emit external signals. Meanwhile, motifs in protein networks represent the structural corrections among internal patterns. The motifs in social networks depict the different connected entities such as persons or companies in online social media. Additionally, in traffic networks, cars and stations are represented as nodes. Also, in labeled networks, the nodes may be labeled by more than one label from a finite set C , whose elements have semantic meanings in the real world. We can represent various labels by using different colors. Therefore, the definition of colored-motifs[5] is shown in Definition 3:

Definition 3. *A motif is a multi-set of elements from the set C of colors.*

Besides the definitions of motif mentioned above, maximal motif[10] is proposed, which means the motif where edges and nodes cannot be added. The maximal-motif does not contain other subgraphs as motifs, thus it cannot diminish the total number of discovered motifs. When we compare with randomly generated networks, three constraints can be defined for classifying the motifs, that is, a probability threshold P, a uniqueness threshold U, and a minimum difference threshold D.

1. A probability threshold P is defined to confine the probability that the frequency of a certain motif appears on a randomly generated network is greater than that of a real network. The threshold P is determined by comparing

it with a large number of randomly generated networks, ensuring that the motif is over-represented in the original network. P can be estimated by assuming a random null hypothesis and $z-scores$, for a given network G_k the definition of $z-score(G_k)$ is described as in the below equation:

$$z-score(G_k) = \frac{f_{ori}}{\overline{f}_{rand}} std(f_{rand})$$

Herein, f_{ori} is the frequency of a given motif in G_k, while $\overline{f}_{rand}$ is the frequency in a randomized network. Also, the $std(f_{rand})$ is the standard deviation of motif frequency in different randomized networks.

2. The threshold U defines a minimum value for significance quantitatively. The second constraint ensures that the motif frequency in the original network, f_{ori} is higher than a uniqueness threshold U.
3. The last constraint is to prevent that the motif is too similar with randomized networks, namely, f_{ori} and $\overline{f}_{rand}$ as well as a narrow distribution in random networks. Thus, f_{ori} should be far larger than $\overline{f}_{rand}$. The equation $f_{ori} - \overline{f}_{rand} > D \times \overline{f}_{rand}$ represents this constraint. In this equation, D is the threshold used to be the limitation of the minimum differences between f_{ori} and $\overline{f}_{rand}$.

3 Related works

3.1 A Motif-Based Approach to Network Epidemics[6]

Background This paper presents an ODE-based method for deriving approximates for epidemics on a network incorporating SIS equations at a triple level. ODE means Ordinary Differential Equation, where the function you want to obtain in the differential equation depends on only one variable. It also presents the application of this approach to a small world network, and compare the application with stochastic simulation.

Method

1. **First order** Define [A] as the number of network nodes in disease state A and [A–B] for the number of pairs of nodes with disease states A and B. Then, define a notation for the total numbers of nodes and pairs as below:

$$[\bullet] := \sum_{A} [A] = N, \qquad [-] := \sum_{A,B} [A\text{–}B] = nN.$$

Fig. 1: First order notation

$$[\wedge] := \sum_{A,B,C} [A\text{–}B\text{–}C], \qquad [\triangle] := \sum_{A,B,C} [\underline{A\text{–}B\text{–}C}].$$

Fig. 2: Second order notation

2. **Second order** Approximation can be improved to the actual behaviour of an epidemic on a network by considering the change of pairs as time passes by. This require the concept of triples, which can be either closed (triangles) or unclosed (lines). The prevalences of these are written as figure 2:
3. **Third order** Time evolution of triples through simple closure approximations has been discussed with the conclusion that it can be preferable to develop more sophisticated pair-level models compared to simple triple-level approaches. Here, consider an approach to triple dynamics that makes use of the full range of possible fourth-order network structures and so avoids the shortcomings of simpler models. While there are only two connected graphs of degree three, there are six connected graphs of degree four, which we write as below:

$$[\nwarrow] := \sum_{A,B,C,D} [A\text{–}\overline{B\text{–}C\ \ D}],$$

$$[\sqcap] := \sum_{A,B,C,D} [A\text{–}B\text{–}C\text{–}D],$$

$$[\leftthreetimes] := \sum_{A,B,C,D} [\underline{A\text{–}B\text{–}C}\text{–}D],$$

$$[\square] := \sum_{A,B,C,D} [\underline{A\text{–}B\text{–}C\text{–}D}],$$

$$[\boxslash] := \sum_{A,B,C,D} [\underline{A\text{–}\overline{B\text{–}C}\text{–}D}],$$

$$[\boxtimes] := \sum_{A,B,C,D} [\underline{\underline{A\text{–}B\text{–}C}\text{–}D}].$$

Fig. 3: Third order notation

Evaluation For faster epidemics, improvements in capturing the average behaviour of a stochastic system is most marked when considering the early growth and transient behaviour of the system, while for slower epidemics the improvement is greatest when considering the endemic state. This is due to the fact that during early growth of quickly spreading infection in a clustered system, **bottlenecking** significantly reduces the incidence rate of infection compared to the mean-field, whereas for slower epidemics it is a long-term behaviour that is

more strongly influenced by clustering, and these facts are captured at increasing levels of detail by both pair- and triple-based approaches.

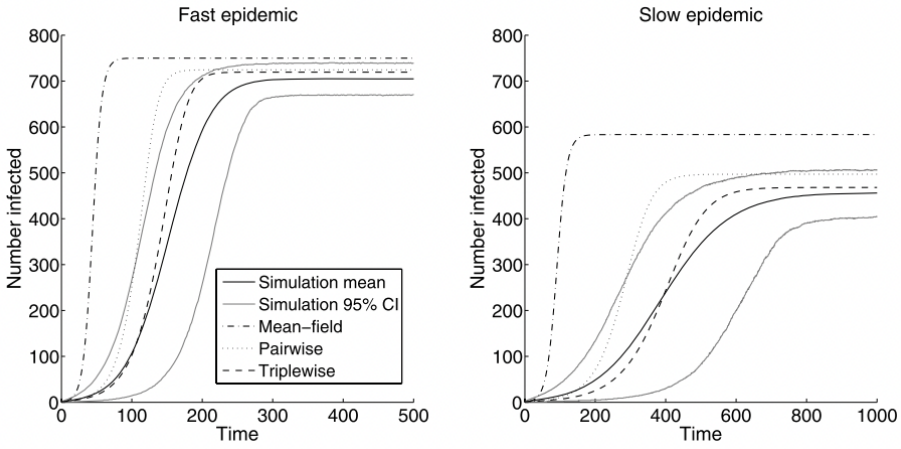

Fig. 4: An SIS epidemic with varying transmissibility on a small world network.

The total number of nodes is $N = 1000$ and the initial number of infectious individuals is 1. The base network is a one-dimensional lattice of degree 2, with randomisation parameter $p = 0.1$. The disease parameters are $\tau = 0.05$,$g = 0.05$ for the 'Fast epidemic' and $\tau = 0.3$,$g = 0.05$ for the 'Slow epidemic'. For the simulations, we show mean and 95% confidence intervals (defined as the range of numbers infected within which 95% of simulations sit at a given time) for 104 runs.

3.2 Identifying critical higher-order interactions in complex networks[7]

Background Hypergraphs are used to model higher-order interactions in complex systems and hyperedges represents the higher-order interactions among entities.

Modeling diffusion in hypergraphs through Laplacians, a key idea in calculating centrality, is not a simple task. This is because hyperedges can include more than two vertices, and edge incidence and vertex adjacency are represented as sets in hypergraphs. To handle this issue, researchers limit their attention to hypergraphs where hyperedges have the same cardinality. However, in this kind of modeling, we have to assume that information only diffuses between fixed size hyperedges. This is not appropriate for hypergraphs since hyperedge can affect other hyperedges regardless of their sizes. Hence, there is a need for more general model for diffusion to detect the critical higher-order interactions.

To overcome the above limitations, this report propose new hypergraph Laplacians based on the diffusion framework that can find the influential higher-order interactions in a hypergraph of any size. The previously developed hypergraph Laplacians are only defined for special hypergraphs where cardinality of hyperedges are limited, which neglects the relations between hyperedges. The model proposed in this paper can complete relations between hyperedges of any size. Also, this paper extended graph centrality measures, namely betweenness (H_{Btw}) and closeness (H_{Cls}) to hypergraphs and rank higher-order interactions based on these measures.

Method We define the graph Laplacian L as $L = D - A$, where D is the weighted degree matrix and A is the weighted adjacency matrix. The graph Laplacian only uses pairwise interactions, i.e., edges, between vertices and ignores higher-order interactions. In this work, diffusion is modeled over a hypergraph inspiring from the the simplicial Laplacians defined in [8]. In the simplicial Laplacians, a hyperedge of size $k+1$ is called a k-simplex. We can define *incidencematrix* as Figure 5:

$$D_p(i,j) = \begin{cases} 1 & \textit{if } \sigma_j^p \textit{ is on the boundary of } \sigma_i^{p+1} \\ 0 & \textit{otherwise} \end{cases}$$

Fig. 5: Incidence matrix

Then, define Laplacian matrix as below:

$$L_i^{up} = W_i^{-1} D_i^T W_{i+1} D_i,$$

$$L_i^{down} = D_{i-1} W_{i-1}^{-1} D_{i-1}^T W_i$$

Then the i-dimensional Laplacian in both directions is as below:

$$L_i^{both} = L_i^{up} + L_i^{down}$$

[**Graph centrality measures**]

$$H_{Btw} = \sum_{s,t \neq u} \frac{n_{st}(u)}{N_{st}}$$

$$H_{Cls}(u) = \frac{1}{C_u}$$

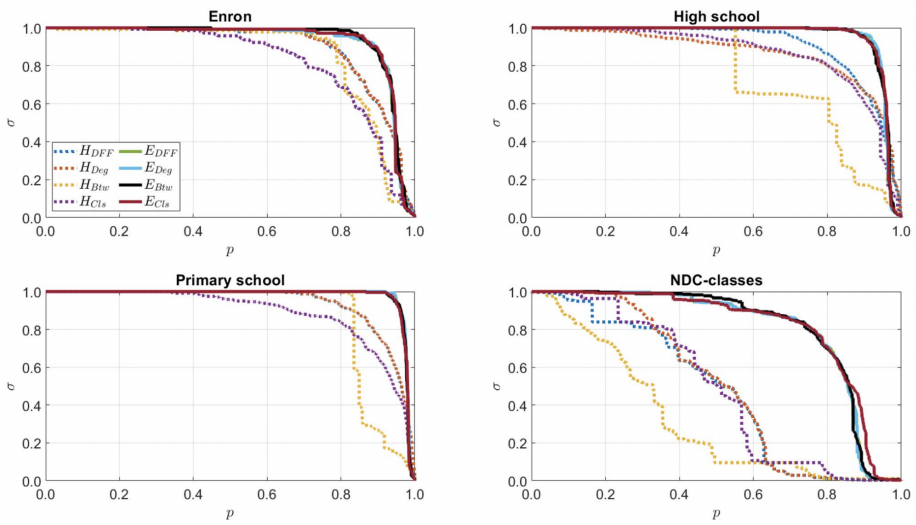

Fig. 6: The size of giant component σ over different ratio p.

Evaluation As we see in Fig 6, the influential higher-order interactions have all faster falls than the influential edges. This means the influential higher-order interactions break down the networks more quickly, due to the fact that whenever we remove an edge between two vertices, these vertices are likely to be still connected with other edges. On the other hand, when we remove a higher-order interaction, it is less likely to have connections between its vertices. We can note that graph measurements H_{Cls} and H_{Btw} curves in Enron and Primary school, and High school and NDC-classes networks respectively fall the faster.

As seen in fig 7 the infection rate is kept constant while the influential hyperedges are obtained by using each of the centrality measures. A higher diffusion index (R_s) determines the effectiveness of each of these methods. It can be concluded that for the Enron data set, degree, betweenness, and closeness centralities is considered to be almost equally effective due to a consistently high R_s value with the increasing ratio of hyperedges. For the High school data set, it can be observed that DFF centrality is the most effective of the four centrality measures with betweenness and closeness centralities overlapping and following the same set of values for Rs with an increase in the varying ratios. In contrast, the degree centrality proved to be least effective for its low R_s value below 0.5. Primary school data set has the most consistent increase in the value of diffusion index R_s for all the centrality measures as there are little or no anomalies within the curves plotted for the four different centrality measures. The degree and the DFF centralities are considered the most effective centralities as they both yield the highest value of R_s in comparison to betweenness and closeness centralities. For the NDC-classes data set, it can be seen that initially the DFF centrality had a higher diffusion index(R_s) in comparison to the degree centrality. Still,

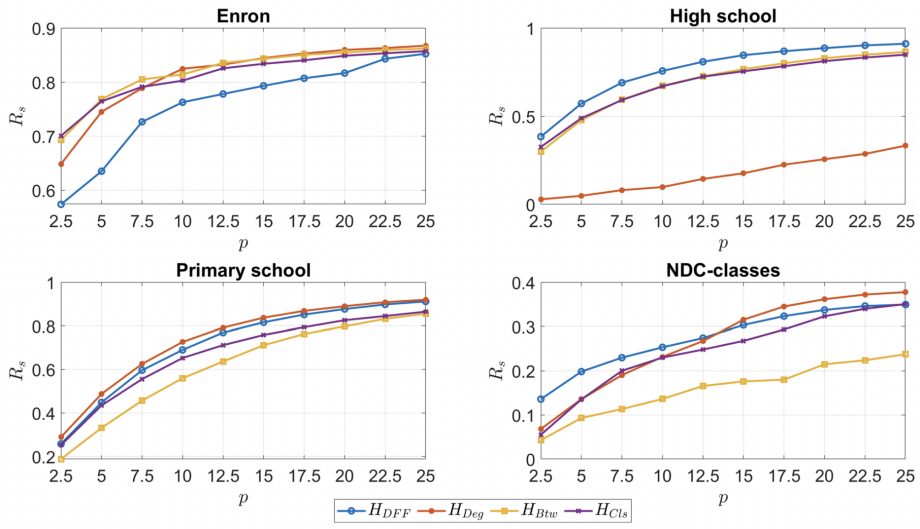

Fig. 7: Varying ratio of hyperedges in the implementation of the SIR Model.

after the ratio (p) increases above 12.5, the degree centrality proves to be more effective than the DFF centrality and the other two centrality measures. In conclusion, the four centrality measures proved to be effective with varying ratios of higher-order interactions used within this SIR model.

4 Questions for further research

As shown in previous sections, we can conclude that in order to grasp the structure of graphs through motifs, we must find the relation between different kind of motifs and the level of contagion proceeded. Below are several research questions that can be derived from the research above:

1. **spatial** At the beginning of a graph with a high rate of infection, what kind of motif appears the most?
2. **temporal** As the degree of infection progresses, what kind of motif appears the most?
3. **overlap** Can Motif determine the degree of overlap or clustering?

5 Conclusion

By looking through the definition for motif and the research related with using motifs as a way to read the patterns of ongoing contagion on networks, we could ask further questions for research that would help to grasp the relation between finding different kind of motif and the progress of epidemics. It would

also be interesting to research more on incorporating motif-based network structure analysis to graph neural networks, where the recognition of patterns could improve the performance of predictions through gnns.

References

1. Ahmed, Nesreen K., et al. "Graphlet decomposition: Framework, algorithms, and applications." Knowledge and Information Systems 50.3 (2017): 689-722.
2. Milo, Ron, et al. "Network motifs: simple building blocks of complex networks." Science 298.5594 (2002): 824-827.
3. Milo, Ron, et al. "Superfamilies of evolved and designed networks." Science 303.5663 (2004): 1538-1542.
4. Soufiani, Hossein Azari, and Edo Airoldi. "Graphlet decomposition of a weighted network." Artificial Intelligence and Statistics. PMLR, 2012.
5. Lacroix, Vincent, Cristina G. Fernandes, and Marie-France Sagot. "Motif search in graphs: application to metabolic networks." IEEE/ACM transactions on computational biology and bioinformatics 3.4 (2006): 360-368.
6. House, Thomas, et al. "A motif-based approach to network epidemics." Bulletin of Mathematical Biology 71.7 (2009): 1693-1706.
7. Aktas, Mehmet Emin, et al. "Identifying critical higher-order interactions in complex networks." Scientific reports 11.1 (2021): 1-11.
8. Horak, D. Jost, J. Spectra of combinatorial laplace operators on simplicial complexes. Adv. Math. 244, 303–336 (2013).
9. Bodó, Ágnes, Gyula Y. Katona, and Péter L. Simon. "SIS epidemic propagation on hypergraphs." Bulletin of mathematical biology 78.4 (2016): 713-735.
10. Parida, Laxmi. "Discovering topological motifs using a compact notation." Journal of Computational Biology 14.3 (2007): 300-323.