A Survey of Graph Data Mining in the Field of Infectious Diseases

Andrew Floyd, Missouri University of Science & Technology

afzm4@umsystem.edu

1 Introduction

As the field of data mining becomes more prevalent and widespread, more models and applications are being developed to help predict and analyze the subject matter of infectious diseases and outbreaks. When it comes to graph and network-based data mining in particular, we will see throughout this survey that infectious diseases prove to be great applications of graphs and social networks as by definition they spread through contact with other individuals. Hence, graph data mining can be used in many different ways from creating models that actually simulate the spread of diseases, to models that use data to track and find useful information about the diseases themselves. This paper will cover some of the more interesting and innovative graph-based models that have been developed regarding infectious diseases and outbreaks. The focus in this survey is more on the side of the methods involving the graph data mining, not necessarily on the results and actual applications of these models (even though those are discussed as well). The next five sections cover five different models and techniques followed by a short summary of some of the non-graph data mining methods used in the field of infectious diseases. All of the methods discussed in this paper were developed and created by the authors of the individual studies and to find out more information about any of the papers covered, see the references section.

2 Detecting Outbreak Clusters

One of the most interesting applications of graph data mining in terms of infectious diseases is that Cori, Nouvellet, et. all [1], who developed a synthesis approach to detecting outbreaks by identifying clusters with graph data structures. They start their process by defining each data stream as a weighted graph, in which the nodes correspond to the cases themselves, while the edges in between two cases are weighted by the pairwise distance (in that particular data stream) between the two cases. This results in the ‘heavier’ edges indicating that the two cases connected by the edge are unlikely to have infected on another. To then only retain the edges, or connections, that are relevant, the graph is then pruned by removing these ‘heavy’ edges. Heavy edges are defined as edges whose weight exceeds that of a predefined cutoff distance. Selecting a good cutoff is critical in their model for identifying the clusters of related cases within the data set. The authors developed a framework for determining this cutoff based on the expected distance distributions between the amount of observed cases during an outbreak. This practice allows them to use pre-existing data about the disease itself like the mutation rate, the distribution of its serial interval (time in-between symptoms arising in a case and its infector) and more. The cutoff used in their applications of the model depended on these factors and more and is slightly different for different data sets. Multiple data streams/graphs can be created based on different kinds of relationships between cases in different spaces (like temporal, spatial, genetic, etc..) all of which can be used to compute pairwise distances between cases. Once these different graphs are constructed and then pruned as described earlier, they individually represent potential links between cases. They then combine these graphs by merging them via intersection so that

the resulting connected components of the final graph has clusters of cases for all types of data, which is illustrated below.

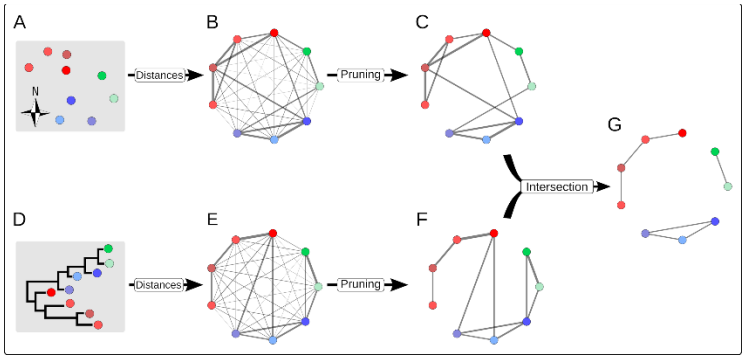

Figure 1: Illustration of combining multiple data streams to identify outbreak clusters

The actual remaining clusters themselves should help identify sets of cases that likely belong to the same 'transmission tree' along with the size of the clusters containing information about the transmissibility of the disease. They are able to use this data to derive estimates of things such as the reproduction number, which is the average number of secondary cases infected by someone who has the disease in a large population. Formally, they define each individual data stream as n, each with their own dissimilarity measure d; hence $d_{i,j}^n$ is the distance between cases i, j with respect to the data stream n. They then define a graph G, where the weight of each edge between nodes i, j is $d_{i,j}^n$. This graph is then pruned by removing all the edges with weights above the predefined cutoff k^n, resulting in a graph $G = \{N, E^n\}$ containing all cases of nodes N that connects cases by edges E^n if and only if the pairwise distance is less then k^n. E^n is formally described by the following:

$$E^n = \{(i,j) | d_{i,j}^n \leq k^n \; \forall \; i,j \in N\}$$

The final step of the approach combines all of the pruned graphs into a single graph G by intersection, which is achieved by retaining the edges present in all of the individual graphs, so that $G = \{N, E\}$ and $E = \{(i,j) | (i,j) \in E^n \; \forall \; n\}$. To put it in layman's terms, the remaining cases that are connected in the final graph have to connected in all of the different data streams so that the connected components of G are able to identify the clusters of cases that are more likely to belong to the same outbreak cluster. How this model is set up allows it to be applied to any number and type of different data streams as really any distance metric can be used for a given data stream. The three main that they tend to focus on are:

1. Data on the timing of infection of cases. The distance is defined as the delay between symptoms of the two cases.
2. Data on the location of cases. Euclidean distance is used between the location of each pair of cases.
3. WGS (Whole Genome Sequencing) of the pathogen sampled from each case. Use the Hamming distance as the distance between two cases, which is defined as the number of Single Nucleotide Polymorphisms (SNPs) between the pair of considered sequences.

All of these of types of data tend to be commonly available which make them good for analysis of many different types of outbreaks. The main testing the authors did on their model involved data from dogs that were infected with rabies reported in Bangui, the capital of the Central African Republic between 2003 and 2012. By outlaying the three sets of data on bar graphs and running analysis on them, they determined that the cutoffs for each of the data streams should correspond to about the 95% quantiles. Once the model was applied to the data streams, three graphs and a final graph were created which can be seen below:

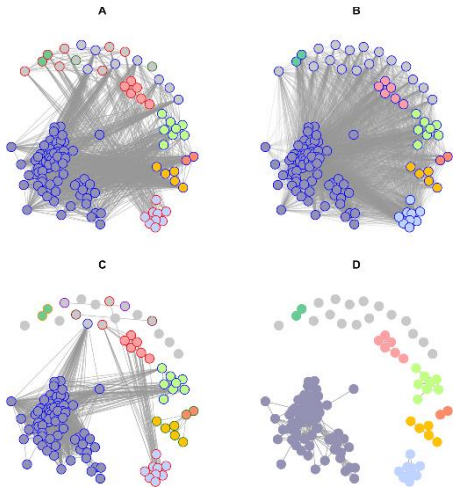

Figure 2: Pruned graphs A (temporal distances), B (spatial distances), C (genetic distances) and D (all three combined)

Using this data, they were able to make predictions on metrics such as the estimated rate of importation of rabies into the population and an estimation of the reproduction number. For more information on the model and the results they received from their trials, see the paper itself [1]. They were able to illustrate that their model could be applied very well to this data on rabies infections within dogs but also discussed how this model would be very useful in real-time for possibly disentangling clusters of related cases from isolated cases (due to separate introductions of the pathogen into the population). This could lead to improvements on control policies, providing real-time information on whether to prioritize different control methods that could reduce transmissions or reduce importations. The generic nature of the model makes it potentially very powerful in terms of applying it to all kinds of infectious diseases, including the current COVID-19 outbreak that is occurring.

3 Identifying High-Risk Individuals

Other studies have been done that look more directly at person-to-person interactions and look to actually create a model where the spread of a disease can be traced through simulation. We see this idea executed in [2], in which the authors use network analysis to an attempt to identify who the high-risk individuals are in a society of people. The goal is to investigate the relation of several measures of network centrality, such as the number of contacts and the measure of the proportion of times an individual lies on a path between other individuals, to the risk of infection after the uprising of a novel infectious disease. They assume that the total population here would start out as being considered fully susceptible to the outbreak and for this purpose that they are living in a somewhat isolated environment (like a particular town or city). Before getting into

how their networks were constructed it is important to note that they generated two types of networks, a 'small-world' network and a random network, the first of which was intended to have a more realistic connection network while the random network is connected in a much more random manner. For both of their network types, the nodes represent individuals themselves or even a group of individuals (such as a family or even a small town) while the edges are contacts by which an infectious agent can travel (such as the transmission of a novel virus by close social contact). It is also assumed that these contacts work in both directions, which makes sense given the definition of an edge a graph, and that these contacts are stable over time. Each network generated consists of 100 nodes, with each node obviously being either connected or not-connected to every other node. A variable, denoted as $\delta_{i,j}$, has the value of 0 if no edge connects nodes i and j while it has the value of 1 if an edge does exist between the two nodes. This is important to understand as the graphs were generated with the criteria that the $\sum \delta_{i,j} = 2000$, that is that the total number of edges in each graph is 1,000 (since each edge connects two nodes by definition). How these edges were selected differed between the two types of networks; for the small-world network each node is assigned into one of 24 random groups with each group (g) having anywhere from 1-8 nodes in it. Nodes i and j were connected with probability 1 if they are in the same group, i.e. $g_i = g_j$, and with a probability of 0.05 if they are not in the same group, $g_i \neq g_j$. This process is repeated until $\sum \delta_{i,j} = 2000$ essentially creating a network that generates clusters of contacts that you would normally see in a real-world situation where individuals form groups of people around them that they tend to associate with. The resulting 'small-world' network has short-path length and high clustering when compared to the random network, making it more characteristic of a real small-world network. The random network was created by randomly selecting two nodes and connecting them with probability 1 assuming they are not already connected by an edge. This was repeated until $\sum \delta_{i,j} = 2000$ and random network was generated where each pair of nodes has the same probability of being connected (regardless of the initial group distribution).

The transmission of the actual infectious agent was done through a simulation model that was applied to both types of networks, starting with the infection phase (S->I). Once again, the model assumes that before the introduction of the infectious agent that the entire population of nodes is susceptible but also does not have the disease already. The outbreak is initiated with the random selection of a node to will be infected. Then for each subsequent time step, if the node n_i has k_i infectious neighbors during the time period, then the node could become infected with the probability of $1 - (1 - \beta)^{ki}$, where β is the probability of transmission per contact. The other half of the simulation is the recovery phase (I->R). Infected nodes were given recovery times of 14 time steps, after which they could no longer transmit the disease along with gaining 'lifelong' immunity for the infection (meaning they could not be infected again). This simulation was done in time steps until the agent had become extinct. At this time the status of each node was recorded (either still susceptible or recovered) along with how long it took each node that was infected to be infected. For each type of network, 3 sets of 500 simulations were completed for their analysis with the duration of recovery being constant at 14 time steps while the probability of transmission (β) varied for the 3 sets; 0.00375, 0.0075 and 0.015.

After running the simulations, the probability of infection for each node was calculated using the number of the simulations in each set in which the node was infected and the total number of simulations. The authors were then able to explore the relation between this probability of infection and some of the characteristics of the node such as its degree, shortest-

path betweenness and random-walk betweenness. They also used generalized additive models to investigate the relationship between different centrality measures and the probability of infection. Besides probability of infection, survival analysis was done by comparing the times to infection with different estimates of survival. The relationship between different continuous covariates and the time to infection were also studied using regression models. Without diving too much into the results of the study, it was found that the relationship between different centrality measures and the probability of infection was similar in the 'small-world' networks and random networks, suggesting that these may be used somewhat interchangeably across different types of networks. Meanwhile, the rate of spread of the agents within the populations and final size of the outbreak were heavily influenced by the network structure. They theorized that the random network had a larger proportion of nodes be exposed to the disease agent then the small-world network while the final size of the outbreaks tended to be smaller in the small-world networks. While the authors admit that the simplicity of their model does cause limitations in regard to the generalizability of the results, they tend to propose that transmission tends to occur more rapidly in small-world networks despite the final outbreak size being smaller. Regarding the centrality measures they tested (random-walk betweenness, shortest-path betweenness, farness), they tended to be correlated with each other along with node degree and proved to be at least as good as other network parameters in predicting the risk of infection. For more information on the model, the simulation results, and their implications see [2].

4 Correlating and Predicting Diseases

The more we tend to look at different methods of using graph data mining in the field of diseases and outbreaks, it becomes clear that there are basically two general types of analysis that can be done. The first of which is constructing models and networks and the running of some sort of analysis to better understand the network and past data entered into it. The second type would be the usage of graph data mining to predict future results based on past data. This next paper [3] takes a look at both of these structures as the authors looked to first build disease networks and then study their structural properties in order to better understand the relationships between them, then secondly created a model a generalized predictive model that takes in data in the form of medical history of a patient. They then extract patient networks from this model that is based on nearest neighbors while also providing a ranked list of other conditions that the patient could be likely to experience in the future. While this model does not necessarily only apply to infectious diseases, it does give us insight into the relationships between different diseases.

To start out, the authors looked at trying to find connections between different diseases. They decided to focus on two similar concepts to try and find these connections, morbidity (which is the number of cases in a given population) and co-morbidity (the co-occurrence of two diseases in the same patient). Using these two concepts, a system to create networks was constructed by linking together diseases A, B that are co-morbid in any one patient while also assigning weights to the edges defined by the following:

$$weight(A,B) = \frac{Co - Morbidity(A,B)}{Morbidity(A) + Morbidity(B)}$$

This formula tends to give higher weights to edges that connect diseases that occur together more frequently due to the numerator while the denominator tends to scale back the weights for

diseases that are highly prevalent in the general population. This ensures that the connections between less-common diseases are brought to the forefront, which is the primary goal of this analysis. The also prune the number of edges present in their network (where the nodes are the diseases themselves and the edges are defined as above) by removing all edges with weight under 0.01, thus eliminating edges between diseases where one is very common and the other is relatively rare as those connections do not provide valuable information. After running this process through their dataset of disease data from Harvard University Medical School, the resulting network had 714 nodes and 3605 edges with an average degree of 10.1. Looking at the visualization of the network below, it was observed that there exists a dense core of nodes, which when looking at the diseases themselves indicated that the core was mostly made up of circulatory and digestive diseases. There are also several tight-knit communities on the outside of the graph which roughly correspond to some of the other major categories of diseases (like accidental injuries, respiratory system, genitourinary diseases, etc.). The one interesting relationship found was that different types of cancers do not tend to form a community with each other, but rather are mixed in with other diseases of the affected organs of that particular type of cancer. The paper itself discusses the results in more detail.

Figure 3: Disease network with edges pruned to weights above 0.01

The other model that was developed in this paper involves using a nearest neighbor network to predict what diseases a particular individual is likely to develop in the future. This is done based on a limited medical history and finding patients that are similar to the patient given, before then giving a risk score to each disease the person has not yet had. A nearest neighbor network is basically a hybrid between two main parts, a near neighbor classifier and a collaborative filtering algorithm that select the k most similar entities to a target and then uses a weighted voting scheme to make predictions on that target. This is defined by the following: each patient P has a set of diseases in their medical history ($diseases(P)$). Then a Jaccard Coefficient, to compute similarity normalized by the total number of diseases two patients share, is computed for each pair of patients P, Q: $s_{Jaccard}(P, Q) = \frac{|diseases(P) \cap diseases(Q)|}{|diseases(P) \cup diseases(Q)|}$. This is the main similarity method used in their work. In traditional nearest neighbor classification, the k most similar other patients would be used to make a prediction for some disease D. But in this case, their model chooses to construct a network by finding the k nearest neighbors of each patient and then connecting them using directed edges. When looking at each disease D, if an immediate neighbor does not have D, a recursive query is done to the nearest neighbor network with depth-first search up to depth l. An example with search depth $l = 2$ is given below:

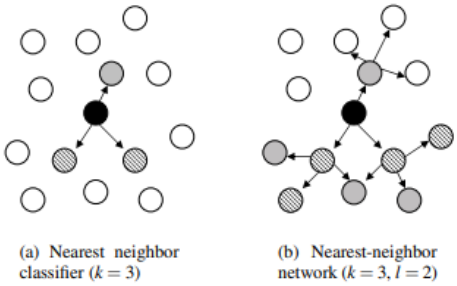

Figure 4: Black nodes indicate the target patient. Neighbors who have the probe disease are shaded, those who do not are striped.

For their testing of this predictive model, they selected a subset of 10,000 patients who had at least five visits to a medical facility. They constructed a network with $k = 25$, hence finding the 25 nearest neighbors. They then iterated over all possible diseases probing the network for each up to depth $l = 3$. A neighbor contributes to the final score proportional to its similarity if he has had the disease, and none otherwise. To see the full results of their analysis, the paper should be referenced [3], but essentially they ranked each disease according to their risk score and then searched this ordered list for diseases that the patient actually developed in future visits (these future visits were not used in the initial computation). They then noted whether they moved up, down or remained in the same position relative to the baseline population (i.e. you want as many as possible moving up). They found on baseline analysis that about 42% moved up. This number was increased by partitioning by gender, up to 46% (due to certain diseases being gender-specific. While additional work is needed to refine this model and minimize the number of diseases missed and maximize the ranking of correct diseases, the ability to accurately predict future diseases is one that could have huge impacts on preventative health care.

5 Mining the Web and Social Media

Another way to approach looking at data from infectious diseases and outbreaks instead of looking at raw data related to disease statistics themselves is to look at how the public reacts to the said outbreak. Especially in today's modern society, social media and online news sources are a major part of how we consume information and form our opinions. It is also important to realize that we can extract and use data from these information outlets, and in this case, create models to help find trends in disease outbreaks. A group of authors looked at how this could be done using specifically web and social media back in 2009 [4]. They hypothesized that these web and social media (WSM) communities play vital roles in any public health response to significant outbreaks, hence it would be possible to leverage data from these sources to not only identify communities but also facilitate outbreak-event detection. And by using graph-based algorithms, this can be done by searching for anomalies in the link-structure within the WSM. While there had been some work done in discovering concepts in linear and attribute-value databases regarding WSM, these authors developed some data mining techniques to discover patterns consisting of complex relationships between entities. In this paper [4], they both discuss a text mining approach used to identify trends in flu posts that correlate to real-world flu-like illness data and a graph-based mining technique (as discussed before) to detect anomalies among flu blogs. We will focus on the second, graph-based approach in this survey paper.

To understand how this graph-based mining algorithm works, we first need to understand what type of data it uses. The authors choose a WSM indexing service called Sprinn3r that conducts real-time indexing of all blogs, processing over 100,000 new blogs per hour. The metadata that is produced from this service includes blog title, URL, post title, date posted, description, full HTML encoded content and more. In this case, the author selected data from an arbitrary period of time (October 5, 2008 – March 21, 2009) and selected English language WSM items that match to the term's 'influenza' and 'flu' anywhere in the content. The main goal of this model was to detect anomalies, which are defined as a surprising or unusual occurrence within a set of data. This paper uses a popular algorithm called Subdue, which more of which can be found in its original paper [5], which basically examines an entire graph and then reports unusual or frequently found substructures within it. Subdue also considers the regularity of the data to determine how likely it is for a substructure to occur given the predictability of the structural data surrounding the substructure. We will discuss the basics of Subdue below to help understand how this algorithm works in context of the WSM postings that are discussed in this paper.

Subdue accepts both undirected and directed graphs as possible inputs but it does require that both the nodes and edges are labeled. It then outputs a set of graphs representing some discovered pattern or concept. This is formally defined as the following: requires a labeled graph $G = (V, E, L)$ as input and output, $V = \{v_1, v_2, v_3, \dots, v_n\}$ (a set of vertices/nodes), $E = \{(v_i, v_j) | v_i, v_j \in V\}$ (a set of edges) and L is a set of labels for both the nodes and edges. Subdue is unsupervised as it searches for a subgraph within the input graph G that best compresses G. These substructures consist of a subgraph definition and all of its occurrences in the input graph. Subdue uses a polynomial-time beam search as its main discovery algorithm, which starts with a set of substructures each with unique node labels, and then applies an operator called ExtendSubstructure to each subgraph in the current state (this algorithm is laid out in pseudocode below). It then extends a subgraph in every possible way by a single edge and a node or only by a single edge (if both of the nodes already exist in the subgraph). These new subgraphs are ordered based value (also known as compression) using Minimum Description Length. This leaves the top subgraphs left for further/more expansion. Once the algorithm terminates, Subdue returns the list of the best subtrees for which the initial graph can be compressed using the best of which. Compression in this case means replacing all instances of the subgraph in the input graph with a single node that represents the subgraph as a whole. Subdue of course can then be applied to this compressed graph again if wanted. This diagram below shows how this algorithm works with a visual example as well.

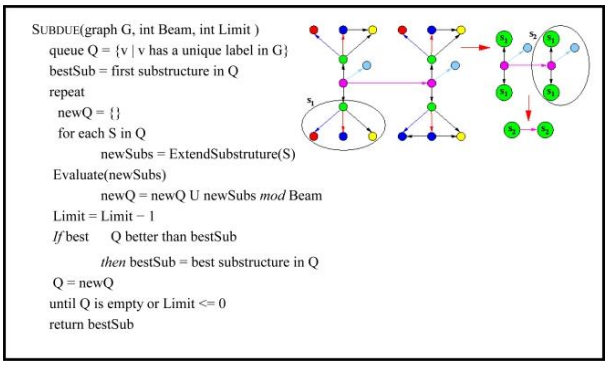

Figure 5: Subdue discovery algorithm and an example

This is the main method that the authors of this paper use to mine the data from WSM influenza based posts. First, they must translate the data found using Sprinn3r into a graph-like structure. They decided to aggregate multiple posts by the same author, hence each representing a unique individual blogger, along with counting multiple tags that go in and out of links only once per blogger. To help enrich the graph, they also connected blogger URLs and tags to nodes that are labeled by the publisher type (for example blog, social media profile, forum, mainstream media, etc.). While this is a bit difficult to visualize from the description, below is example graph that the authors gave which is helpful to understand how the graph structure works.

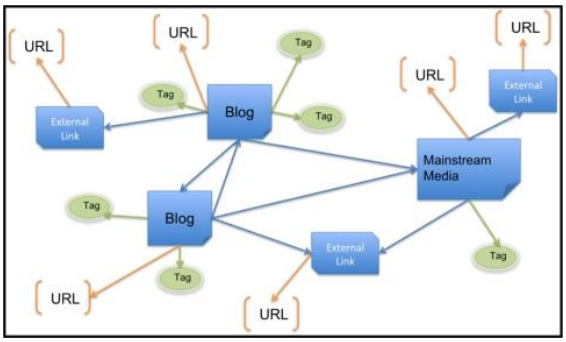

Figure 6: Example representation of a graph used by Subdue

As seen above, essentially URLs are removed from their main WSM 'parent' creating a link between the URL and WSM article itself. This was done to allow the Subdue algorithm to find subgraphs from blogs with differing content, like news and personal blogs, in addition to the traditional URL structures. The results from this analysis were interesting as they were able to formulate a list of subgraphs found by Subdue over the different time periods that the data was collected while also identifying if the subgraph corresponds to an anomaly. These anomalies would therefore be used for the purpose of disease outbreak detection. The first main anomaly of interest found in their testing was from a United Kingdom Yahoo Answers site, which directly corresponds to the time frame in which the UK was having one of its worst flu seasons in years. They also found a high correlation between personal blog sites that mention the flu (through Subdue identifying novel subgraphs of course). Another interesting anomaly found that showed a larger number of subgraphs occurrences was made up of MySpace posts that discusses different American Idol contestants that contracted the flu and were unable to perform as well during the show. This model shows very interesting potential to be applied to all kinds of web and social media and could be used to detect trends and anomalies regarding both disease outbreak and really any other topic of interest.

6 Simulating Outbreaks using a Network Model

Most of the attempts at creating networks based on human contact, which of course is how highly infectious diseases are usually spread, both covered in this paper and outside of it have mostly been based on data collection such as surveys, other social networks [4], and even mobile cellular data. Instead of using those methods, which each have their pros and cons, our next paper [6] looks infectious diseases could spread through a school environment (high school to be specific). The goal of this paper was to analyze a contact network observed at a typical American high school by using a SEIR (susceptible, exposed, infectious and recovered) simulation model

to investigate the spread of influenza. The authors also looked to implement and test various immunization strategies to evaluate their efficacy in reducing the spread of infectious diseases within a school setting. For this survey paper, we will mostly focus on the network and simulation aspect of the research. And while this method isn't necessarily using data mining techniques, it represents an important area of research in terms of utilizing graph networks to help prevent and predict infectious diseases.

Before creating a network, first a dataset was needed. The data the authors collected covered the closer proximity interactions (CPIs) of 94% of a particular high school population including over 650 students and over 130 staff and other people containing 2,148,991 unique CPI records. Adding in other factors as well, the aggregate network for this data set consisted of a weighted undirected graph where the nodes represent individuals in the school and the edges represent contacts, that are also weighted for the duration of the contact between two individuals. While traditional models tend to assume that all individuals have around the same number of contacts. This network model tries to remedy this problem by accounting for the fact that the majority of the contacts made were found to be relatively short. This was done by recalculating all the statistics of the network with a minimum requirement for contact duration, c_m, meaning all edges with weight $< c_m$ are removed from the graph. While the network model itself is fairly simple, the authors did notice that it exhibited typical 'small-world' properties such as a high transitivity (clustering coefficient) and a short average path length for all values of c_m. They also found the community structure to be very high, increasingly so with higher values of c_m, indicating that more intense and longer contact durations tend to occur more often in subgroups and less often between different subgroups.

The simulation itself was done using an SEIR simulation model parameterized with data from previous influenza outbreaks. Transmission in the model occurs only among the individuals in the school as each individual/node is classified as either susceptible, exposed, infectious or recovered. Assuming no vaccination, all nodes are initially classified as susceptible and at a random time step during the first week of the simulation one node is chosen as the 'index' case and his status is set to exposed. Each time step represented 12 hours along with being broken up into days with the requirement that transmission can only occur during the day and on weekdays (to follow the idea of an actual high school). Transmission of the disease from an infectious to a susceptible node occurred at a probability of 0.003 per 20 seconds of contact (this value could be changed but was chosen based on another study done on infectious outbreaks). This means that the probability of transmission per time step (12 h) from an infectious individual to a susceptible individual is $1 - (1 - 0.003)^w$, where w is the weight of the contact edge. On infection, an individual will move into the exposed class (meaning they have been infected but not infectious). After the incubation period, an exposed individual will become symptomatic and move into the infectious class. On the day that a node becomes infectious, the duration of all contacts of that node is decreased by 75%. In the following days all contacts between that node and others is decreased by 100% until recovery (simulating an individual staying at home). Once someone is infectious, recovery occurs with a probability of $1 - 0.95^t$ per time step (t is the number of time steps spent in the infectious state). After twelve days in the infectious state, the node will recover if the recovery has not already happened. The simulation stops when the number of both exposed and infectious nodes is back at zero. Vaccination strategies were also tested and added to the simulation that yielded different results obviously based on when and how the vaccine was distributed to the school population. Without going into great details regarding the results of this

network model and corresponding simulation, the authors did find that their outcomes were very similar to that of absentee data from the same school during the second wave of H1N1 in the fall of 2009. They also found a strong correlation between the size of a given outbreak caused by index case node i and the strength of the node (weighted equivalent of the degree) representing that node. The correlation was less between the outbreak size and the actual degree of the index node. For more information on this model, its results and the effects of vaccination on the simulation, please see [6].

7 Other Data Mining Applications within Infectious Diseases

While the majority of this survey paper focuses on graph and network related data mining models that have been applied to the field of infectious diseases, there has also been tons of research done using other forms of data mining. We will review a few of those here (in less detail then that of the graph data mining models).

Social media is a popular way to accumulated data and has been found to be quite useful in terms of tracking infectious diseases. These authors of this paper proposed an unsupervised machine learning model that they proved has the ability to identify real-world latent infectious diseases [7]. They achieve this goals by mining social media data from sources such as Twitter, using messages from the platform with both user and temporal information. The thoughts of users from their tweets regarding symptoms, pain and body parts are all identified from the data mined. Then symptoms weighting vectors for each individual and time period are created and infectious disease related information is retrieved from these vectors. They found that they were able to identify latent infectious diseases, without prior information, quicker than when the disease is formalized by public health officials.

Another approach that covers a multitude of different topics is introduced as the ‘Framework for Infectious Disease Analysis’ [8]. This proposed model is a software environment and conceptual architecture that covers fields such as data integration, data visualization, situational awareness, prediction and intervention assessment. As complex as it sounds, this is a much more complete framework for predicting and tracking diseases as this model collects bio-surveillance data using natural language processing. They use both unstructured and structured data from all kinds of sources and then apply machine learning concepts, such as linear regression and decision tree-based boosting, before using multi-modeling to analyze different disease dynamics. Data such as websites, social media and news feeds are all used to extract information for these uses. For the disease predictions themselves, classification machine learning models are used including random forests, boosting and support vector machines.

8 Conclusion

The importance and impact that modeling, predicting and forecasting infectious diseases has on the medical community and society as a whole is massive. As discussed throughout this paper, there is a multitude of research being done in the field of graph data mining relating to infectious diseases and outbreaks. Different approaches to modeling diseases using network-like layouts can be seen as social media, health care data and even general population information has been used to create these models. As the impact of infectious diseases on our population is so huge,

especially with the recent COVID-19 outbreak that has crippled society like no other infectious disease before has, the models and methods such as those covered in this survey have to potential to have life-altering implications.

9 References

[1] Cori A, Nouvellet P, Garske T, Bourhy H, Nakouné E, et al. (2018) A graph-based evidence synthesis approach to detecting outbreak clusters: An application to dog rabies. *PLOS Computational Biology 14(12), e1006554.*

[2] R. M. Christley, G. L. Pinchbeck, R. G. Bowers, D. Clancy, N. P. French, R. Bennett, J. Turner. Infection in Social Networks: Using Network Analysis to Identify High-Risk Individuals. *American Journal of Epidemiology, Volume 162, Issue 10, 15 November 2005:* 1024–1031.

[3] Karstn Steinhaeuser and Nitesh V. Chawla. A Network-Based Approach to Understanding and Predicting Diseases. *In Social Computing and Behavioral Modeling, December 2008.*

[4] Corley, Courtney D et al. "Text and structural data mining of influenza mentions in Web and social media." *International journal of environmental research and public health vol. 7,2, 2010*: 596-615.

[5] Cook DJ, Holder LB. Substructure discovery using minimum description length and background knowledge. *J. Artif. Int. Res. 1993.* 1:231–255.

[6] Salathé, Marcel & Kazandjieva, Maria & Lee, Jung Woo & Levis, Philip & Feldman, Marcus & Jones, James. "A High-Resolution Human Contact Network for Infectious Disease Transmission". *Proceedings of the National Academy of Sciences of the United States of America, 107, 22020-5, 2010.*

[7] Sunghoon Lim, Conrad S. Tucker, Soundar Kumara. "An unsupervised machine learning model for discovering latent infectious diseases using social media data." *Journal of Biomedical Informatics, Volume 66, 2017*: 82-94.

[8] Erraguntla, M., Zapletal, J., & Lawley, M. "Framework for Infectious Disease Analysis: A comprehensive and integrative multi-modeling approach to disease prediction and management." *Health Informatics Journal, 25(4), 2019:* 1170–1187.