Network Analysis and Recommendation for Infectious Disease Clinical Trial Research

Magdalyn Elkin
Dept. of Biology, Florida Atlantic University
Boca Raton, FL 33431, USA
melkin2017@fau.edu

Whitney A. Andrews
Dept. of Computer Science & Eng., Florida Atlantic University
Boca Raton, FL 33431, USA
andrewsw2014@fau.edu

Xingquan Zhu
Dept. of Computer Science & Eng., Florida Atlantic University
Boca Raton, FL 33431, USA
xzhu3@fau.edu

ABSTRACT

Clinical trials are crucial for the advancement of treatment and knowledge within the medical community. Since 2007, US federal government took the initiative and requires organizations sponsoring clinical trials with at least one site in the United States to submit information on these clinical trials to the ClinicalTrials.gov database, resulting in a rich source of information for clinical trial research. Nevertheless, only a handful of analytic studies have been carried out to understand this valuable data source. In this study, we propose to use network analysis to understand infectious disease clinical trial research. Our goal is to answer two important questions: (1) what are the concentrations and characteristics of infectious disease clinical trail research? and (2) how to accurately predict what type of clinical trials a sponsor (or an investigator) is interested in? The answers to the first question provide effective ways to summarize clinical trial research related to particular disease(s), and the answers to the second question help match clinical trial sponsors and investigators for information recommendation. By using 4,228 clinical trails as the test bed, our study involves 4,864 sponsors and 1,879 research areas characterized by Medical Subject Heading (MeSH) keywords. We extract a set of network measures to show patterns of infectious disease clinical trials, and design a new community based link prediction approach to predict sponsors' interests, with significant improvement compared to baselines. This trans-formative study concludes that using network analysis can tremendously help the understanding of clinical trial research for effective summarization, characterization, and prediction.

KEYWORDS

Clinical trial, Infectious disease, Network analysis, Community, Link prediction

ACM Reference Format:
Magdalyn Elkin, Whitney A. Andrews, and Xingquan Zhu. 2019. Network Analysis and Recommendation for Infectious Disease Clinical Trial Research. In *Proceedings of The 10th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, Niagara Falls, NY, USA, Sept. 7-10, 2019 (BCB'19).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

BCB'19, Sept. 2019, NY, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Clinical trials carry out tests on human participants *w.r.t.* different interventions, including new medications or treatment, in order to understand and answer meaningful clinical questions [1]. These studies, usually made through joint efforts between pharmaceutical companies and research institutions, are critical for discovering new treatments that are potentially more effective than current solutions to diagnose, treat, and reduce the risk of disease.

Publicly available information of clinical trial studies is valuable for researchers, because it helps them understand what is effective in treatment and what is not. Such studies are also helpful for researchers and doctors to decide if the side effects of a new treatment are acceptable when weighed against the benefits offered by the new treatment [2]. Efforts to regulate documentation on clinical trials help circulate pertinent results among the research community to stimulate further advances in a quicker period of time.

Despite of the vital importance, for years, obtaining clinical trial documentations is a daunting task. Prior to 2007, only the publication results from FDA-approved drugs were mandated in the U.S. In fact, research from that time period had shown that unsuccessful clinical trials were far less likely to be published [3]. This negatively impacts the research community because new research groups would not know that a particular clinical trial had already been executed and that the results were unsuccessful, thus costing time and money. As such, a mandated singular repository for such trials is extremely beneficial, because it is a useful resource for researchers to learn from a singular pool of documented successes and unsuccessful ventures for clinical trials [4], and it also helps researchers understand the current state of the pharmaceutical sector [5].

1.1 ClinicalTrials.gov Initiative

In 2007, the U.S. Federal Government took the initiative and issued regulations surrounding clinical trials being conducted in the U.S. and other clinical trials under the command of U.S. Food and Drug Administration (FDA). These newly mandated regulations required the submission of new clinical trial information to an already existing database: ClinicalTrials.gov. Since 2007, the mandated regulation initiative requires organizations sponsoring clinical trials with at least one site in the United States to submit information on these clinical trials to the ClinicalTrials.gov database. The number of registered studies since the mandate in 2007 has increased by 256,123 (from a 12 year time gap, that is approximately 21,000 registered per year on average). Prior to the mandate, from 2000 to 2007, the

difference of registered clinical trials was 46,174. Even though this is a seven year time gap, only approximately 6,500 were registered per year on average. The difference between the two time periods is evident. ClinicalTrials.gov currently lists 304,654 studies since May 2019 with locations in all 50 states and in 208 countries [4]. Even though ClinicalTrials.gov is an abundant source of clinical trial study with longest history and largest complete data [5], it is, unfortunately, an underutilized information source for health industry and life science research [3], considering the rapid growth of the field and a rather limited number of studies being made based on CLinicalTraisl.gov reports. The amount of registered clinical trials in the U.S. is exponentially increasing, and according to a recent report by Grand View Research, global clinical trial market is expected to reach 65.2 billion dollars by 2025 [6]. The global clinical trial market is expected to grow at a compound annual growth rate of over 5.5% from 2017 to 2025. North America was found to dominate the overall market in terms of revenue sharing in 2015 attributed from increased research and development in the region as well as the presence of big outsourcing firms. More interestingly, the growing prevalence of disease and incidence of new disease is expected to give further boost to the clinical trial market. Such a growing trend naturally raises questions on how to better analyze and utilize existing clinical reports to benefit industry, academia, and individuals [7].

1.2 Network Analysis of Clinical Trial Study

Although clinical trial reports provide rich information for analysis, health and biomedical research are known to contain complex objects and relationships. Since 2007, researchers have already begun to conduct investigations in to understanding trends in clinical research by utilizing ClinicalTrials.gov database. Through understanding clinical trial distribution by medical condition, it is possible to anticipate future important medical developments and find where the innovation maybe first adopted [3]. Despite of rather limited research efforts, the rapid growth of clinical trial studies results in many new medical conditions, new medical issues, and diseases appearing over time. Determining the relationship and expected projection of the trend of clinical trials is difficult when there are a large variables from different sub-domains. In this paper, we will use network analysis to investigate distributions, correlation, and predicted connections of infectious disease related clinical reports in relationship to sponsors/investigators.

Network analytics are commonly used to understand structure, development, and relationships of complex systems. Such analysis provides valuable information about the systems, such as link prediction, correlation, or degree distribution [8]. For example, a network based analysis has been used to study long-term collaboration in pharmaceutical industry [5].

Different from existing research, in this paper, we propose to use bipartite network [8] to represent clinical trial research entities and analyze their relationships. We will create networks to represent infectious disease clinical trials, and understand characteristics of such networks, including most commonly studied areas and community structures of infectious disease clinical trials. After that, we will propose a community based link prediction (CLP) to predict links for information recommendation.

2 RELATED WORK

There has been work already done on network robustness an developing local and global clustering coefficients in respect to two-mode networks. Two-mode networks are also known as affiliation or bipartite networks. Robustness is defined as the ability of a network to continue performing well when it is subject to failures or attacks [9]. Traditionally, to analyze a two-mode network, they are converted in to a one- mode network, however, in doing so, a lot of structural information could be lost that is pertinent. A paper by Opshal [10], proposed a redefined global clustering coefficient defined by a four path. It was noted that the proposed global clustering coefficient did have a limitations in that the primary node mus be the first and last node of the 4-path. For our purposes this was not an issue.

Other work has been done in determining network structure and distance in bipartite graphs, such as the small world study. In the paper by Robins, they compared empirical bipartite graphs to stimulated random graph distributions. The bipartite graph is examined directly as the previous study did, as opposed to converting the graph into two 1-mode graphs. This paper introduced the reinforcement coefficient to evaluate model robustness that measures for localized bipartite cycles. They found that two networks may share many similarities and some differences.

Both of the aforementioned papers did cover innovative network analysis measures that were executed on different data sets for examination (they did not use CinicalTrial.gov). In the following subsections, we review studies that have direct correlation to the data set used in this study or similar data sets in structure (one data set is sponsored by the World Health Organization (WHO) for international clinical trials that is not enforced).

2.1 Data Analytics of Clinical Trial Reports

A handful of studies have been previously made to understand CLinicalTrails.gov database using data analytics [3, 7, 11]. One study evaluates the wholeness and effectiveness of ClinicalTrials.gov database, and overviews the importance of the source for describing information about the landscape for clinical trials. This study found that the mandated variables, labeled by the authors for examination sake, are potentially valuable as a research source. In fact, the incompleteness of the data was found to be less that 3%. Few databases meet this level of completion. The participants are required to ensure the data are correct in this database, which aids in the completeness of that data, therefore proving its high quality and importance as a research tool. Although this tool did not directly use any network analytic to examine any behavioral patterns, they authors did use an XML parser that compiled an SQL database with all current information at the time, therefore analyzing the content and quality of the data [3]. This study reaffirms the confidence in our choice for database selection and its importance in the field for research and general field evaluation.

Another study conducted in 2015 [11] also analyzes data in registered clinical trials, but uses an international database that is not mandated. The database used is the International Clinical Trials Registry Platform that is supported and created by the World Health Organization (WHO) to monitor various countries in their pharmaceutical endeavours as well as the various global corporations and sponsors. The goal of the study was to analysis the different trends

in registered clinical trials over a span of nine years based on clinical developments and their respective causes. The study found that by analyzing the differences in those trends between countries and regions, the increase in trial registration had different trajectories in different parts of the world. This study, however, is partially incomplete because their is no global coalescence nor local measures to enforce the clinical trials to be reported to this database. This study shows the dominance of the ClinicalTrials.gov database in the quality of the data and completeness, thus yielding results that would fairly project industry trends.

2.2 Network Analysis of Clinical Trial Reports

Network analysis has been previously used for disease-drug bipartite network, which is close to what we will be studying [8]. In their study, a link prediction method in the disease-drug bipartite network is proposed. The method is called the internal links method with the base being similarity based link prediction. The data was obtained from www.drugs.com. The results showed that the proposed internal link method had a good percentage of success than the other similarity based link prediction methods [8]. In our research, we chose to link prediction using collaborative based filtering based on the community based structure.

A recent study conducted in 2018 [5] uses ClinicalTrials.gov database to analyze long-term collaboration. The purpose was to test the clinical trials information for observing the status of the collaboration network and open innovation in the pharmaceutical industry. The authors focused on a time period from 1980 to 2017 in increments of 10 years. The breakdown of the creative network was based on types on agencies participating in the clinical trials in the collaboration network. The study showed several different understandings of the relationships among the listed pharmaceutical companies, research institutes, and universities, and their mechanisms. While the number of clinical trials among agencies has stagnated since the 2000s, the number of collaborations continue to grow. The leading entities in current clinical trials are different from the intermediators establishing many partnerships on the clinical trial collaboration network [5]. Even though this study did look at pharmaceutical agencies an collaborating entities, it has not examined the behavior between these entities and specific diseases being treated through clinical trials. In fact, it is noted at the beginning of the paper that nor many cases have analyzed the clinical trials database, even though it is the information source with the widest coverage for the pharmaceutical industry [5]. This study was published very recently, thus showing the lack of research in this particular field and that there is much room for further study.

3 DATA

In this study, we download 4,228 infectious disease clinical trial reports from ClinicalTrials.gov database as our test bed. The downloaded reports include past, current, and future clinical trials during 1991-2023. An example of the report (encoded in XML format) is shown in Figure 1.

Because the main goal of our research is to understand characteristics of infectious disease clinical trials (*e.g.* what are the main diseases studied in the clinical trials, who are interested in infectious disease, and what are other areas they are interested in), we extract investigators/sponsors and clinical trial areas from two XML tags: (1) investigator information: ⟨overall_official⟩, and (2) area of clinical trials: Medical subject headings (MeSH) ⟨mesh_term⟩). An investigator is the individual (*e.g.* a physician or a researcher) who submits and is in charge of the underlying clinical trial. Research areas are Medical Subject Headings (MeSH) Terms which roughly define the focused research topics the underlying clinical trial. MeSH was created by the US National Library of Medicine as a method to describe a wide variety of biomedical topics to properly index articles in MEDLINE [12]. In this study, the research area was determined by intervention and condition MeSH words from the file. A clinical trial report often contains one or multiple sponsors/investigators, and multiple research areas.

Formally, we use s to denote a sponsor/investigator and use k to denote a keyword of research area. Likewise, we use $\mathbb{S}$ to denote the set of all sponsors, and $\mathbb{K}$ represents the set of all keywords (research areas).

In this research, we collected 4,288 infection disease related clinical trial reports, from which we extracted 4,864 investigators (*i.e.* $|\mathbb{S}|$=4,864) and 1,879 research areas (*i.e.* $|\mathbb{K}|$=1,879).

```
<clinical_study>
  <!-- This xml conforms to an XML Schema at:
    https://clinicaltrials.gov/ct2/html/images/info/public.xsd -->
  <brief_title>Total Marrow Irradiation; Autologous Stem Cell Transplantation for Relapsed or Refractory
  <brief_summary>
    <textblock>
      The investigators hypothesize that conformal radiation will allow the administration of
      higher doses of external beam radiation to marrow based malignancies than total body
      irradiation (TBI)without increasing the toxicity to normal tissues beyond that induced by TBI.
    </textblock>
  </brief_summary>
  <overall_official>
      <last_name>Harold L Atkins, MD</last_name>
      <role>Principal Investigator</role>
      <affiliation>Ottawa Hospital Research Institute</affiliation>
  </overall_official>
  <condition_browse>
      <!-- CAUTION:  The following MeSH terms are assigned with an imperfect algorithm          -->
      <mesh_term>Multiple Myeloma</mesh_term>
      <mesh_term>Neoplasms, Plasma Cell</mesh_term>
  </condition_browse>
    <!-- Results have not yet been posted for this study                                        -->
  </clinical_study>
```

Figure 1: An example of a clinical trial report encoded in XML format. The file contains important information of the clinical trial, such as project title, summary, investigators/sponsors carrying out the clinical trial, keywords *etc.* In this paper, we extract investigator information from the ⟨overall_official⟩, and extract Medical subject headings (MeSH) ⟨mesh_term⟩ as the area of the clinical trial research.

4 METHODS

In this section, we first introduce the bipartite network used to model clinical trial sponsor-area relation. Then we propose to use community detection to find group of investigators sharing similar research topics. After that, we propose a community based link prediction (CLP) to recommend research areas for investigators.

4.1 Bipartite Graph for Clinical Trial Sponsor-Area Relationship Modeling

Because clinical trials involve complex sponsors and research area relationships, *e.g.* a sponsor may be interested in multiple closely related (or interdisciplinary) research areas and results from one

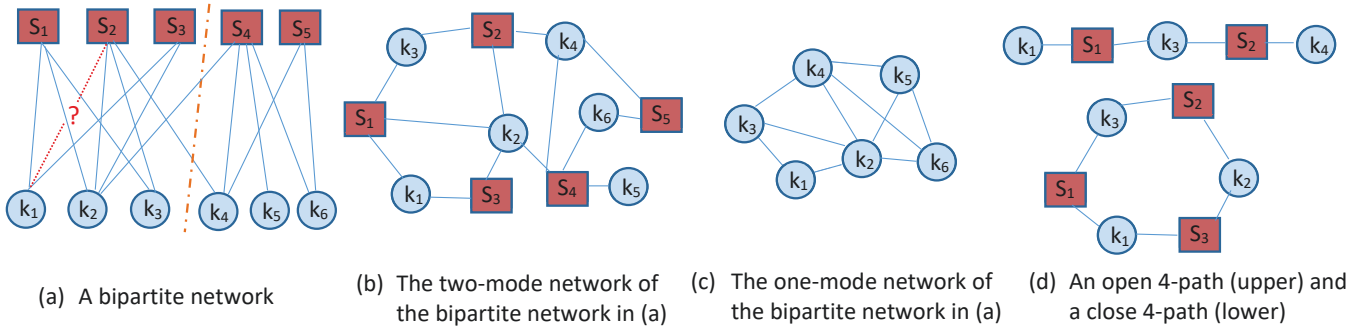

Figure 2: An conceptual view of using bipartite graph for clinical trial sponsor-area relationship modeling. (a) shows a bipartite network where upper pink squares denote sponsors and lower blue circles indicate research areas. A blue solid line denotes an edge, indicating that a sponsor has conducted a clinical trial on the connected area. The brown dot-dash line separates the networks into communities suggesting that sponsors and their research areas fall into two groups. The red-dash line (with a question mark) is the predicted link, predicting that s_2 is interested in k_1 (although the connection currently does not exist); (b) shows the two-mode network of the bipartite network in (a); (c) shows one-mode network which omits sponsor nodes in the bipartite graph. Two area nodes are connected if they both connect to one sponsor node in the bipartite network in (a); and (d) shows examples of close 4-path (lower) and open 4-path. A close 4-path in (d) is a circle in the one-mode network in (c).

research area may be beneficial to another areas, properly modele sponsor-area relationship has immediate benefits to both researchers, industry, and participants. Meanwhile, the nature of pair-wise sponsor and research area bound provides a bipartite relationship for analysis. So we use bipartite network as the underlying data structure to support our analysis.

Formally, a bipartite network $\mathbb{G} = \mathbb{G}(\mathbb{V}, \mathbb{E}, \mathbb{W})$ is a graph where the node set, $\mathbb{V}$, can be partitioned into two disjointed sets ($\mathbb{V}$=$\mathbb{V}_1 \cup \mathbb{V}_2$). No node belongs to both sets of $\mathbb{G}$, ($\mathbb{V}_1 \cap \mathbb{V}_2$). Edges, $\mathbb{E}$ connect from one node set to the second node set ($\mathbb{E} \subset \mathbb{V}_1 \times \mathbb{V}_2$). Sponsors represent one set of nodes and research areas represent the second set of nodes. An example bipartite network is shown in Figure 2(a). The degree of a node, deg(v), is the number of edges incident to node v. In an undirected bipartite graph, the deg(s) is the number of k nodes that s is connected to and vice versa. In Figure 2(a), deg(s_1)=3.

If a clinical trial had multiple sponsors, edges are created from all investigators to research areas. The weight $\omega(e)$ of each edge represents the number of times an investigator being connected to a research area. In the case that an investigator name does not exist in the clinical trial report, the trial's sponsor was used instead. For simplicity, we will refer to investigators and sponsors as sponsors. To decrease the sparsity of the network, MeSH words that contain a comma were separated into two research areas, *e.g*, "Influenza, Human" was separated into "Influenza" and "Human."

To explore the bipartite network and determine the most popular research areas, research area nodes degree and PageRank scores are calculated. PageRank ($\boldsymbol{r}$) was calculated using the eigenvector formulation.

$$\boldsymbol{r} = \boldsymbol{M}\boldsymbol{r} \tag{1}$$

Where $\boldsymbol{M}$ is a transition matrix of the network, denoting $\frac{1}{|N_{out}(j)|}$ if there is an edge between node j to node i, and 0 otherwise. PageRank, $\boldsymbol{r}$, corresponds to the principle eigenvector of $\boldsymbol{M}$.

4.2 Clinical Trial Network Community Detection

Community detection aims to find connected groups of nodes within a network. In Figure 2 (a) the dot dash line represents the split of the bipartite network into two communities such that $\mathbb{C}_1$ contains node set $\{s_1, s_2, s_3, k_1, k_2, k_3\}$. And $\mathbb{C}_2$ contains node set $\{s_4, s_5, k_4, k_5, k_6\}$. Network community detection was done using the LPAwb+ algorithm created by Beckett. [13] Communities are found by distinct modules that consists of a combination of two node types in a weighted bipartite network. The algorithm computes modules based on two stages. The first stage sets a label, g_x, for each node based on maximizing the modularity score for a weighted bipartite network, Q_W, defined in Eq. (2) [13] [14].

$$\begin{aligned} Q_W &= 1M \sum_{u=1}^{r} \sum_{v=1}^{c} (\tilde{W}_{uv} - \tilde{E}_{uv}) \delta(g, h_v) \\ &= 1M \sum_{u=1}^{r} \sum_{v=1}^{c} \Big(\tilde{W}_{uv} - \frac{y_x z_v}{M} \Big) \delta(g, h_v) \end{aligned} \tag{2}$$

Where g and h are node types, sponsors and research areas, g_u is a sponsor node and h_v is a research area node. The Kronecker delta function $\delta(g_u, h_v)$ equals one when nodes u and v are in the same module, or community, or zero otherwise. $\tilde{E}$ is a matrix of no interactions between two nodes, $\tilde{W}$ is the weighted incidence matrix, y is the incidence matrix row totals and z is the column totals. The node's label, g_x, is found maximizing equation (3). In the first stage sponsor nodes are updated using information from research area nodes and research area nodes are updated using information from sponsor nodes. Labels are updated until modularity score, Q_W, can no longer increase [13]

$$\begin{aligned} g_x &= \Big(\sum_{v=1}^{c} (\tilde{W}_{xv} - \frac{y_x z_v}{M}) \Big) \delta(g, h_v) \\ &= \Big(\sum_{v=1}^{c} (\tilde{W}_{xv} \delta(g, h_v) - \sum_{v=1}^{c} (\tilde{W}_{xv} - (y_x z_v m) \delta(g, h_v) \Big) \end{aligned} \tag{3}$$

In the second stage, groups of communities are merged together. Each module consists of nodes that share the same label. Communities are merged if merging increases network modularity. This is repeated until it isn't possible to increase network modularity by merging any more communities [13]. Each community, $\mathbb{C}_c$, contains a distinct subset of s and k such that $\mathbb{V} = \mathbb{C}_1 \cap \mathbb{C}_2$.

Since Infectious Disease Clinical Trials cover many diverse research areas, it is important to determine the robustness of communities. Robustness can be defined as the ability of a network to withstand failures [9]. The transitivity of social networks has widely been studied [15] [10]. Transitivity can define connectivity in a network by defining the number of connections between connected nodes. It is measured by the fraction of connected triangles to the number of connected triplets [15]. A triangle is where V_1 and V_2 are connected and are both connected to V_3. A connected triplet is where V_1 is connected to V_2, and V_2 is connected to V_3 and there is no connection between V_1 and V_3. To measure transitivity, the clustering coefficient, C_c, is often used [15] [10]

$$C_c = \frac{3 \times (\#\ of\ triangles)}{\#\ of\ connected\ triplets} \tag{4}$$

This is frequently used in one-mode networks, an example of one-mode network is shown in Figure 2(c). A high clustering coefficient indicates high robustness. If a graph is completely connected, *e.g*, all nodes connect to each other, $C_c = 1$. If the graph has no triangles, $C_c = 0$.

However, the global clustering coefficient can't be applied to two-mode networks, such as a bipartite network (Figure 2(a)). By definition in a two-mode network, nodes in set $\mathbb{S}$ only connect to nodes in set $\mathbb{K}$, thus a triangle will never form. [10] So to determine robustness, we used two coefficients created for bipartite two-mode networks. The first is a global coefficient, GC_c, which measures the number of closed 4-paths compared to the number of 4-paths. A path is a sequence of connected distinct nodes. An open 4-path is one where the first and last node do not connect. In Figure 2(d) (upper panel) nodes $\{k_1, s_1, k_3, s_2, k_4\}$are on an open 4-path. A closed 4-path (also called a 4-cycle) is a path where the first and last node connect. In a bipartite graph, they are connected by a 5th node. In Figure 2(d) (lower panel) nodes $\{k_1, s_1, k_3, s_2, k_2\}$are on a closed 4-path, closed by s_3. A 4-cycle is the smallest cycle possible in a two-mode network. GC_c=1 if all 4-paths in a bipartite network are closed, and 0 if all 4-paths are open [10]

$$GC_c = \frac{\#\ of\ closed\ 4\text{-}paths}{\#\ of\ 4\text{-}paths} \tag{5}$$

The second measure was the reinforcement coefficient, RC_c, which measures the number of closed 3-paths compared to total 3-paths in the network. It's considered reinforcement between two sponsors rather than a measure of clustering between a group of sponsors. A high reinforcement coefficient indicates localized closeness in a bipartite network [16]

$$RC_c = \frac{\#\ of\ closed\ 3\text{-}paths}{\#\ of\ 3\text{-}paths} \tag{6}$$

A community whose research areas only connect to one sponsor, or multiple sponsors only connect to one research area would not have a value for either GC_c or RC_c coefficient (an example is shown in Figure 7). In this case, we consider this type of community as an invalid community.

4.3 CLP: Community Based Link Prediction for Clinical Trial Research Recommendation

In order to accurately recommend/predict research areas interesting to a sponsor, we propose to use link prediction to find connections between sponsor nodes s and research area node k that currently do not exist. In Figure 2(a) the red dashed-line with a question mark is a predicted link that suggests that node s_2 is interested in node k_1.

Link Prediction has been extensively studied in research and many methods [17], such as similarity-based, supervised learning based, or collaborative filtering based approach, have been used for link prediction. In the following, we first discuss existing collaborative filtering based link prediction, and then propose our community based link prediction.

4.4 GLP: Global Link Prediction using Collaborative Filtering

User-based collaborative filtering [18] is generally performed to predict the votes of a user_a on a particular item_j by comparing user_a to other users in the dataset (user_i) who have a vote on item_j ($v_{i,j}$). The vote for item_j by user_a ($P_{a,j}$) is determined by Eq. (7). In this study we are predicting weight of linkage between a sponsor and a research area. The highest predicted weight would indicate that research area is interesting to the sponsor (*e.g.* the topic he/she may be interested in pursuing in the future). For clinical trial bipartite network, we treat users as sponsor nodes (s) and items as research area nodes (k), and the vote indicates the weight value between sponsor node and research area node. The highest value $P_{s,k}$ for k would indicate the top one predicted research area and so on.

$$P_{s,k} = \bar{v}_s + \kappa \sum_{i=1}^{n} \omega(s,i)(v_{i,k} - \bar{v}_i) \tag{7}$$

Where n is the number of sponsors, $\omega(s,i)$ denotes the similarity between two sponsors s and i, $v_{i,k}$ denotes the weight value (vote) between sponsor i and research area k, and κ is a normalization parameter. $\bar{v}_i$ is the average weights of sponsor i, which is defined in Eq. (8) ($\mathcal{N}_i$ denotes the set of research area nodes connecting to sponsor i) [18].

$$\bar{v}_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} v_{i,j} \tag{8}$$

In summary, $P_{s,k}$ denotes sponsor s's weight on research area k, and $P_{s,k}$ is the average weights of sponsor s plus the weighted summation of all other sponsors' weight on research area k. The more similar two sponsor nodes are, the more similar their weights for research area k will be.

In this study, we used cosine similarity to measure similarity between two sponsors. Assume $\boldsymbol{A}$ and $\boldsymbol{B}$ are vector representation for the sponsor of interest ($\boldsymbol{A}$) and sponsor to compare ($\boldsymbol{B}$) from the weighted incidence matrix, $\tilde{W}$, their similarity is calculated as

follows, where m is the dimension of the vector.

$$\omega(\boldsymbol{A},\boldsymbol{B}) = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{||\boldsymbol{A}||\cdot||\boldsymbol{B}||} = \frac{\sum_{i=1}^{m}\mathbf{A}_i\mathbf{B}_i}{\sqrt{\sum_{i=1}^{m}(\mathbf{A}_i)^2}\sqrt{\sum_{i=1}^{m}(\mathbf{B}_i)^2}} \tag{9}$$

4.5 CLP: Community Based Link Prediction

In previous section, our research has observed that sponsor-area relationship has strong community tie, where sponsors/investigators are very likely to be interested in research areas within the same community. This is mainly because that biomedical research has strong domain requirement, where an investigator trained in one area is often only specialized in limited relevant areas. Meanwhile, as interdisciplinary and cross domain research continuously grows, more and more clinical trials involve team of experts from experts multiple domains, which essentially complicate the community structure in clinical trials.

Motivated by the above observations, we propose to use community based link prediction to recommend links. The detailed process of the CLP is shown in Algorithm 1, which includes two major components: (1) create bipartite network from clinical trial reports; (2) detect community from bipartite networks, and (3) apply user-based link prediction to each community to find links.

5 RESULTS

5.1 Infectious Disease Clinical Trial Network Characteristics

In Figure 3, we report the degree distributions of research area nodes ($\mathbb{K}$) of the infectious disease clinical trial network (in log-log scale). For comparisons, Figure 4 shows the degree distribution of all sponsor nodes ($\mathbb{S}$). Both figures show scale-free degree distributions with long-tail phenomenon, meaning that many sponsors focus on rather few topics (Figure 4) and some common topics receive many attentions by researchers (Figure 3).

From Figure 4, we observed that the maximum deg(s) is 140, and there are 25 sponsor nodes with degree = 140. Interestingly, all 25 sponsor nodes are connected to the same 140 research area nodes. This may indicate that the 25 sponsors worked together on multiple clinical trials.

In order to find areas most commonly investigated in infectious disease clinical trial, Table 1 reports the top 10 research area nodes (k) in the network by node degree, showing that the maximum deg(k) is 864. For comparisons, Table 2 reports the top research areas according to the PageRank scores.

The results from Tables 1 and 2 show that the top 10 research area nodes (k) by degrees and by PageRank scores are very similar, but slightly different. For example, Hepatitis A has the 10th largest degree, of 235, but is not included in the top 10 PageRank score k nodes. PageRank is determined by the importance of the links that point to any particular node. If a k node has a large number of degree, it's possible that the s nodes pointing to it, don't connect to other important research areas or don't have many other connections. Thus a node with large degree may not have a large PageRank score.

Algorithm 1 CLP: Community Based Link Prediction for Clinical Trial Research Recommendation

input: (1) Infectious Disease Clinical Trial Report Dataset: $\mathcal{D}$; (2) Number of recommendations: k
output: Top-k recommended sponsor-area pairs: $\mathcal{SA}_k$
$\mathbb{E} \leftarrow \emptyset$ Initialize edge list
for each clinical trial report $d \in \mathcal{D}$ **do**
  $\mathcal{S} \leftarrow$ Extract sponsors from d. {sponsor nodes}
  $\mathcal{A} \leftarrow$ Extract areas from d. {area nodes}
  $\mathbb{E} \leftarrow \mathbb{E} \cup \{\mathcal{S} \times \mathcal{A}\}$. {sponsor-area edges}
end for
$\mathbb{G} \leftarrow \mathbb{E}$ {Create Network from edge lists}.
repeat
  $Q_W \leftarrow$ Maximizing modularity score of $\mathbb{G}$ using Eq. (2)
until Convergence
for each vertex $x \in V$ **do**
  $g_x \leftarrow$ Find its modularity-based label using Eq. (3)
  $\mathcal{G} \leftarrow \mathcal{G} \cup g_x$
end for
$\mathbb{C} \leftarrow$ Find communities using modularity labels $\mathcal{G}$
for each community $c \in \mathbb{C}$ **do**
  $GC_c \leftarrow$ Find its global coefficient using Eq. (5)
  $RC_c \leftarrow$ Find its reinforcement coefficient using Eq. (6)
  $\mathcal{V} \leftarrow \mathcal{V} \cup c$, if c is a valid community according to its GC_c and RC_c scores
end for
for each valid community $c \in \mathcal{V}$ **do**
  $\mathbb{E}_c \leftarrow$ Find edges directly connected to any vertex in c
  for each sponsor $s \in c$ and a research area k **do**
    if $e_{s,k} \notin \mathbb{E}_c$ {link $e_{s,k}$ does not exist} **then**
      $P_{s,k} \leftarrow$ Find s's scores *w.r.t.* k within the network $\mathbb{E}_c$ using Eq. (7)
    end if
  end for
end for
Rank all sponsor nodes in $\mathcal{V}$ in descending order based on their $P_{s,k}$ sores.
$\mathcal{SA}_k \leftarrow$ top-k nodes on the ranked list
return $\mathcal{SA}_k$.

Table 1: The top 10 research area based on node degrees

Research Area	Degree
Infection	864
HIV Infections	656
Communicable Diseases	637
Tuberculosis	412
Pneumonia	399
Hepatitis	309
Sepsis	295
Malaria	259
Anti-Bacterial Agents	256
Hepatitis A	235

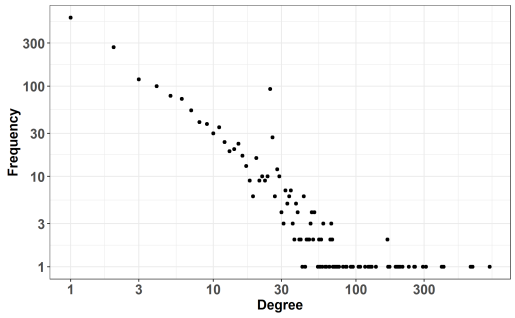

Figure 3: Degree distributions of research area nodes ($\mathbb{K}$) in log-log scale.

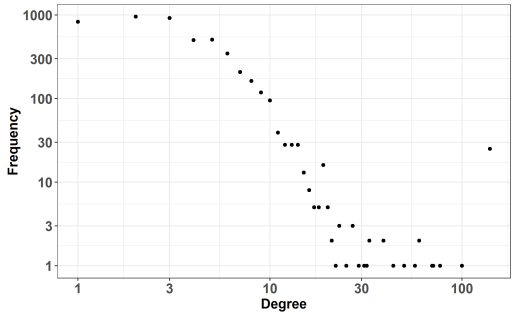

Figure 4: Degree distributions of sponsor nodes ($\mathbb{S}$) in log-log scale.

Table 2: The top 10 research areas based on PageRank scores

Research Area	PageRank Score
Infection	0.016107339
HIV Infections	0.015347930
Tuberculosis	0.012601791
Communicable Diseases	0.011201864
Pneumonia	0.010338744
Malaria	0.006997163
Sepsis	0.006997163
Hepatitis	0.005717856
Human	0.004635713
Influenza	0.004583310

5.2 Infectious Disease Clinical Trial Community Detection Results

Table 3 lists the summary of detected infectious disease clinical trial communities. Overall, we found 478 communities $\mathbb{C}$, and 139 of them have valid GC_c and RC_c scores (these communities are listed as "Valid" in Table 3). In total, all valid communities have 3,662 sponsor nodes (s) (75.29% of all sponsor nodes) and 1,304 research area nodes (k) nodes (69.40% of all research area nodes), indicating that valid communities cover large portions of networks. For all valid communities, their global clustering coefficients, GC_c range from 0.4 to 1 with average of 0.9814, and their reinforcement coefficients, RC_c range from 0.054 to 1 with average of 0.728.

Table 3: Summary of clinical trial community detection results. Each of the six columns represents: (1) valid *vs.* invalid communities, (2) number of detected communities ($|\mathbb{C}|$), (3) number of sponsor nodes ($|s|$), (4) number of research area nodes ($|k|$), (5) the average Global Coefficient (GC_c), and (6) the reinforcement coefficient (RC_c), respectively.

	$\lvert\mathbb{C}\rvert$	$\lvert s\rvert$	$\lvert k\rvert$	GC_c	RC_c
Valid	139	3662	1304	.981	.054
Non-valid	339	1202	575	NA	NA

In order to understand the structure of the detected infectious disease clinical trial communities, Figure 5 shows the structure of the community $\mathbb{C}_3$, which includes nine sponsor nodes (s) and six research area nodes (k). The k nodes for $\mathbb{C}_3$ are also listed in Table 4. For $\mathbb{C}_3$, RC_3 = 0.9310345 and GC_3 = 1. As you can see in Figure 5, the subgraph is almost a complete graph, thus GC_3 = 1. There are two k nodes that only connect to one s node. The sponsor nodes in GC_3 all are connected with each other on four research areas, which could be from the same clinical trial or multiple clinical trials. The remaining two research areas are present from other studies. As such these sponsors have a large reinforcement coefficient, RC_3 = 0.9310345.

When analyzing the community $\mathbb{C}_3$, we can find that there is highly localized clustering within the community. The highly connected research areas of $\mathbb{C}_3$ are, in fact, different types of Penicillins, which are common antibiotics. The two areas with degree = 1 are "Procaine" and "Trepenomal Infections". Looking at the graph, it may not seem as if "Procaine" and "Trepenomal Infections" belong in $\mathbb{C}_3$, however they are both conceptually linked to Penicillin. Treponemal diseases are bacterial infections which can cause syphilis, bejel and yaws. Trepenomal diseases are caused by various *Treponema* bacterial species. Penicillin has been the primary treatement for treponemal disesases for the last 50 years [19]. Procaine is a local anesthetic [20], which has often been mixed with penicillin to create a combination antibiotic and local anesthetic. Procaine penicillin has been used to treat infections due to *Listeriamonocytogenes*, *Treponema* and *Acintomyces* bacterial species among other infections [21]. "Penicillin G Procaine" is another research area in $\mathbb{C}_3$, further validating "Procaine" in this community.

Table 4: Research areas of community $\mathbb{C}_3$

Research Areas	Research Areas
Penicillin G	Penicillin G Benzathin
Penicillin G Procaine	Penicillins
Procaine	Treponemal Infections

The structure of the community $\mathbb{C}_{100}$ is shown in Figure 6, which consists of 35 sponsors (s) and nine research area key words (k). The detailed research areas for $\mathbb{C}_{100}$ is also listed in Table 5. For the community $\mathbb{C}_{100}$, RC_{100} = 0.7139049 and GC_{100} = 0.9699529. In general, as the number of nodes increases, the reinforcement coefficient decreases. This could be due to the fact that with more nodes,

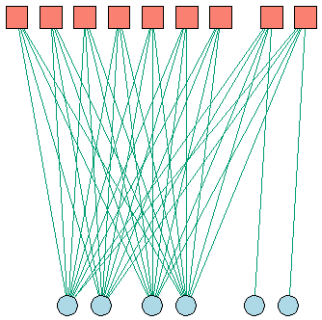

Figure 5: The structure of the community $\mathbb{C}_3$ which consists of 15 nodes (nine sponsor nodes and six research area nodes). The pink squares indicate sponsors and the blue circles indicate research areas (Table 4 lists detailed research areas).

the lower the amount of local clustering within the community. A detailed analysis of the community $\mathbb{C}_{100}$ shows that this community consists of antibiotics and conditions. Amoxicillin and Oxacillin are both broad-spectrum Penicillin-class antibiotics. [22] [23] Clavulanic acid is an additive to amoxicillin with can increase amoxicillin's antibiotic properties [23]. Gentamicin is a broad spectrum aminoglycoside antibiotic, Aminoglycosides are a class of antibiotics that act by creating holes in a bacterial cell's membrane [24]. In $\mathbb{C}_{100}$, only two k nodes, "Oxacillin" and "Intestinal Obstruction", have degree = 1. However it's clear that Oxacillin is related to Amoxacillin. Meanwhile, "Intestinal Obstruction" is conceptually related to "Abdominal Pain". Therefore, this community is consisted of three conditions: "Intestinal Obstruction", "Abdominal Pain", and "Acute Disease", and five interventions that are all possible antibiotic treatments for the three conditions.

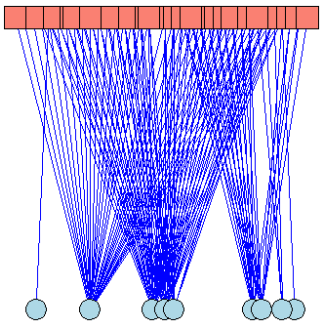

Figure 6: The structure of the community $\mathbb{C}_{100}$ which consists of 36 sponsor nodes and nine research areas. The pink squares represent sponsors and the blue circles denote research areas. (Table 5 lists detailed research areas).

An example of an invalid community, $\mathbb{C}_{12}$, is shown in Figure 7. This community consists of only one sponsor node (s) and three research area nodes (k). In Table 6, we list k nodes of $\mathbb{C}_{12}$. Because there is only one sponsor node, $\mathbb{C}_{12}$ has RC_3 = NA and GC_3 = NA, and degree of k nodes within $\mathbb{C}_{12}$ all equal to one. As a result, it is treated as an invalid community. Our analysis shows that an

Table 5: Research areas of community $\mathbb{C}_{100}$

Research Areas	Research Areas
Abdominal Pain	Acute Disease
Amoxicillin	Clavulanic Acid
Clavulanic Acids	Oxacillin
Intestinal Obstruction	Gentamicins
Amoxicillin-Potassium Clavulanate Combination	

invalid community consists of only one or two sponsors from a single clinical trial.

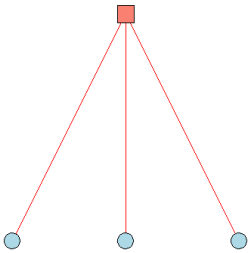

Figure 7: The structure of an invalid community $\mathbb{C}_{12}$ which has one sponsor node and three research area nodes. The research areas are listed in Table 6.

Table 6: List of research areas of community $\mathbb{C}_{12}$

Research Areas	Research Areas
Focal Segmental AIDS-Associated Nephropathy	Glomerulosclerosis

5.3 Infectious Disease Clinical Trial Recommendation Results

In order to validate the performance of the proposed community-based link prediction for clinical trial recommendation, we carry out following designs to remove a small portion of connections from the networks as benchmarks, and then compare different methods' performance to accurately predict these "removed" links.

To create benchmark links for prediction, we generate following three benchmark node sets, representing sponsor nodes with increasing number of connections.

- $\mathbb{S}_{[2,6]}$**:** randomly select 100 sponsor nodes from $\mathbb{S}$ where each selected sponsor has minimum 2 edges and maximum 6 edges. This set represents sponsors with normal degree of connections (majority sponsors belong to this category, as shown in Figure 4).
- $\mathbb{S}_{(6,10]}$**:** randomly select 100 sponsor nodes from $\mathbb{S}$ where each selected sponsor has minimum 7 edges and maximum 10 edges. This set represents sponsor with a high degree of connections.
- $\mathbb{S}_{(10,\infty)}$**:** randomly select 100 sponsor nodes from $\mathbb{S}$ where each selected sponsor has minimum 11 edges. This set represents sponsors with a very high degree of connections.

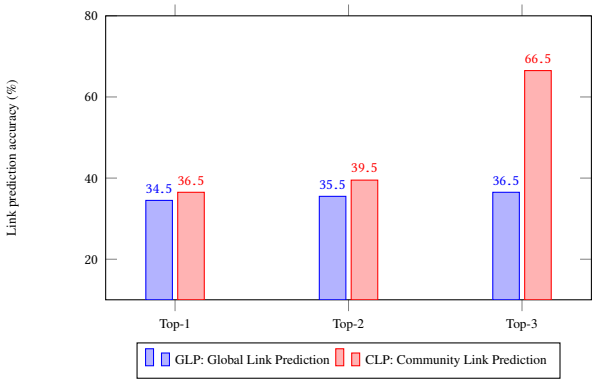

Figure 8: Link prediction accuracy comparison between global link prediction (GLP) and the proposed community link prediction (CLP) on benchmark node set $\mathbb{S}_{[2,6]}$. The $x-$axis denotes the top-k prediction, and the $y-$axis denotes the link prediction accuracy.

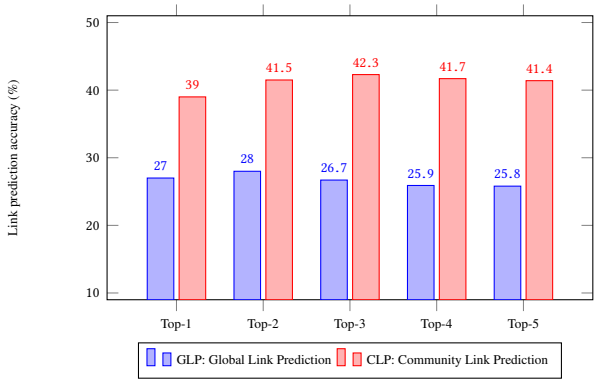

Figure 9: Link prediction accuracy comparison between global link prediction (GLP) and the proposed community link prediction (CLP) on benchmark node set $\mathbb{S}_{(6,10]}$. The $x-$axis denotes the top-k prediction, and the $y-$axis denotes the link prediction accuracy.

After creating the above three benchmark node sets, for each node in any of the selected sets, half of its edges are removed and the removed edges are used as benchmark edge set of the selected node set. If a link predict method predict a research area that was removed, then the prediction is accurate (*i.e.* the predicted result is the one that was removed). By doing so, we know the ground truth of the links and can therefore compare algorithm performance.

Figure 8 reports the results with respect to node set $\mathbb{S}_{[2,6]}$. Because for each selected node, half of links were removed from a subset of 100 s nodes with 2-6 degrees (148 sponsor-area edges are removed in total), we compare each method's top-1, top-2, and top-3 accuracy to validate their performance. For top-1 accuracy, it means that for a sponsor, each algorithm only reports the top-1 recommendation, and calculates the accuracy across all 100 sponsors.

The results from Figure 8 show that for top-1 and top-2 accuracy, GLP and CLP were very similar. GLP obtains 34.5% accuracy for top-1 and 35.5% accuracy for top-2. CLP obtains 36.5% accuracy for top-1 and 39.5% accuracy for top-2. The difference was greatest for top-3, where GLP's accuracy is 36.5%, whereas and CLP's accuracy is 66.5%. This experiment shows that community based link prediction (CLP) is consistently better than global based approach (GLP). For majority sponsor nodes in the network, CLP can accurately predict/recommend its link with at least 36.5% accuracy.

Figure 9 shows the prediction results from benchmark node set $\mathbb{S}_{(6,10]}$ which represents nodes with relatively high degree of connections. In this round, half the links are removed from a subset of 100 s nodes with 6-10 degrees (345 sponsor-area edges are removed in total), therefore, we use top-1,$\cdots$, top-5 accuracy. For top-1 to top-5, GLP obtained 27%, 28%, 26.7%, 25.9% and 25.8% accuracy, respectively. CLP had higher accuracy, for top-1 to top-5, CLP obtained 39%, 41.5%, 42.3%, 41.7% and 41.4% accuracy.

Figure 10 reports the prediction results from benchmark node set $\mathbb{S}_{(10,\infty)}$ which represents nodes with very high degree of connections. In this experiment, sponsor nodes with degree >10 are selected and half of their links are randomly removed. Recall that there are sponsor nodes with up to 140 degree, so some selected sponsor nodes have up to 70 edges being removed. Therefore, we report accuracy from top-1 to top-70 prediction accuracy in Figure 10. Overall, the results show that GLP has an average accuracy of 31.8% and CLP's average accuracy is 40.6%, which is much higher than GLP.

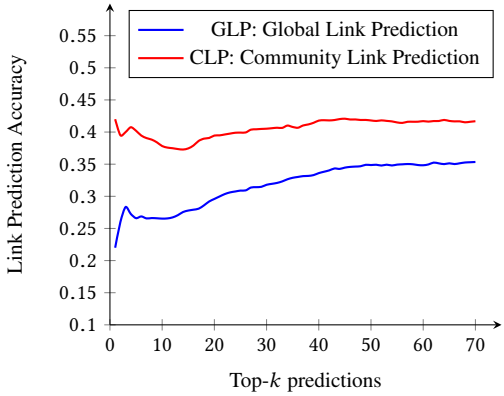

Figure 10: Link prediction accuracy comparison between global link prediction (GLP) and the proposed community link prediction (CLP) on benchmark node set $\mathbb{S}_{(10,\infty)}$. The $x-$axis denotes the top-k prediction, and the $y-$axis denotes the link prediction accuracy.

6 DISCUSSIONS

In summary, our research discovers communities of sponsors and research areas and validates that our CLP method is more accurate than global link predictions.

In our experiments, the main bulk of frequency degree distribution found was between degree 10 and degree 30. The research area with the highest degree found was Infection with a degree of 864 followed by HIV Infection and then communicable diseases, as seen in 1. This coincides with current industry trends. A study done by Goswami et al. in 2013 found HIV/AIDS trials to be the largest

subset of Infectious Disease Clinical trials in the ClinicalTrials.gov database [25]. They found Influenza Vaccine trials to constitute the second largest subset of Infectious Disesase Clinical Trials [25]. In this study, Influenza was not one of the top k nodes by degree, but it was one of the top k nodes by PageRank Score. The degree of a node is denoted as the number of edges that connect to the node. In this study, the edges are based on the connection of a sponsor to a research area in Infectious Disease. This directly related back to the current trends determined by the sponsors on a specific research field. It is interesting to note that these top 10 popular fields are not concentrated within the bulk of the degree distribution frequency, which would mean that there are topics that are less researched by sponsors and yet they appear frequently in terms of the degree.

Communities of sponsors and research areas can help effectively summarize a wide range of research areas and have the potential to bring together sponsors who are connected by only shared research areas. Our research finds communities to summarize research areas and support link predictions. To determine the robustness, we use a modified global coefficient for bipartite networks and a reinforcement coefficient. Robustness would be the ability of a network to continue performing well when it is subject to failures or attacks. As seen in Table 3, the average modified global coefficient is quite high while the reinforcement coefficient is quite low for valid communities. The global coefficient is a more indicative measure of robustness, whereas the reinforcement coefficient is a more localized measure that would indicate the reinforcement of a potential relationship between sponsors within the community. As shown in the example communities, valid communities consists of conceptually linked research areas, whereas invalid communities have research areas that are being studied by only one or two sponsors.

In this study, we randomly remove links from the network to validate our link prediction methods. We found that CLP is more accurate than GLP. It is interesting to find that in some cases, a high RC_c score would lead to more accurate prediction in CLP, but not in GLP. However this is not always the case. It could be that in some cases, some sponsors have research areas that are very localized, compared to the entire network. In these localized communities, the sponsor's connections may not reach outside the community, thus a high RC_c score would indicate that link prediction within that community would be very accurate. But if a researcher has connections outside the community, then a high RC_c score won't indicate the accuracy level for link prediction.

In this study, we left out invalid communities. It is possible that a sponsor may have a potential future link to a research area that is only being studied by one or two sponsors. However, it would be hard to use structural network based analysis alone to determine this. Unpopular research areas would have to be linked to sponsors based on node-content instead of network structure.

7 CONCLUSIONS

In this study, we proposed to study relationships between investigators/sponsors and research areas in infectious disease clinical trials extracted from ClinicalTrials.gov. We argued that ClinicalTrials.gov is a valuable, yet under utilized, data source. By using bipartite graph to create infectious disease networks between sponsors and research areas, our research studied characteristics of the networks, detected communities from the network, and further proposed a community based link prediction to recommend research areas for sponsors. Experiments and validations confirmed that the proposed method is much more accurate in recommending links for infectious disease clinical trial research. The framework proposed in the paper (including network analytics) can be generalised and extended to any other clinical trial areas, such as heart disease.

ACKNOWLEDGEMENT

This research is sponsored by the US National Science Foundation through Grants IIS-1763452 and CNS-1828181.

REFERENCES

[1] L. M. Friedman, C. D. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger, *Fundamentals of Clinical Trials, 5th Edition*. Springer, 2015.
[2] F. S. Collins, "The importance of clinical trials," *NIH MedlinePlus*, 2011.
[3] H. E. Glass, L. M. Glass, and J. J. DiFrancesco, "Clinicaltrails.gov:an underutilized source of research data about the design and conduct of commercial clinincal trials," *Therapeutic Innovtion Regulatory Science*, no. 49(2), pp. 218–224, 2014.
[4] "Trends, charts, and maps," May 2019. [Online]. Available: https://clinicaltrials.gov/ct2/resources/trends#RegisteredStudiesOverTime
[5] H. Yang and H. J. Lee, "Long-term collaboration network based on clinicaltrials.gov database in the pharmaceutical industry," *MDPI Sustainability*, pp. 1–14, January 2018.
[6] "Report: Global clinical trials market is expected to reach 65.2b by 2025." Sep 2017. [Online]. Available: https://www.centerwatch.com/news-online/2017/09/25/report-global-clinical-trials-market-expected-reach-65-2b-2025/
[7] R. M. Califf, D. A. Zarin, J. M. Kramer, R. E. Sherman, L. H. Aberle, and A. Tasneem, "Characteristics of clinical trials registered in clinicaltrials.gov, 2007-2010," *Journal of the American Medical Association*, vol. 307.
[8] E. Gundogan and B. Kaya, "A link prediction approach for drug recommendation in disease-drug bipartite network," *International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–4, September 2017.
[9] W. Ellens and R. Kooij, "Graph measures and network robustness," *arXiv:1311.5064v1*, November 2013.
[10] T. Opsahl, "Triadic closure in two-mode networks: Redefining the global and local clustering coefficients," *Social Networks*, no. 35, July 2011.
[11] R. F. Viergever and L. Keyang, "Trends in global clinical trial registration: an analysis of numbers of registered clinical trials in different parts of the world from 2004 to 2013," *BMJ Open*, pp. 1–15, July 2015.
[12] M. Huang, A. Neveol, and Z. Lu, "Recommending mesh terms for annotating biomedical articles," *Research and Applications*, May 2011.
[13] S. Beckett, "Improved community detection in weighted bipartite networks," *Royal Society Open Science*, no. 3, 2016.
[14] C. Dormann and R. Strauss, "A method for detecting modules in quantitative bipartite networks," *Methods in Ecology and Evolution*, no. 5, pp. 90–98, 2014.
[15] M. Newmann, "Scientific collaboration networks. i.network construction and fundamental results," *Physical Review E*, no. 64, June 2001.
[16] G. Robbins and M. Alexander, "Small worlds among interlocking directors: Network structure and distance in bipartite graphs," *Computational Mathematical Organization Theory*, no. 10, pp. 69–94, 2004.
[17] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," *J. of the American Society for Information Science and Technology*, vol. 58.
[18] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," *Proc. of the Fourteenth conference on Uncertainty in Artificial Intelligence*, pp. 43–53, July 1998.
[19] M. Marks, A. Soloman, and D. Mabey, "Endemic treponemal diseases," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, no. 108,10, pp. 601–607, 2014.
[20] "Procaine." [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/compound/procaine
[21] A. M. Bazakis and A. J. Weir, *Procaine Penicillin.* StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing, 2019.
[22] "Oxacillin." [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/compound/oxacillin
[23] P. Todd and P. Benfield, "Amoxicillin/clavulanic acid. an update of its antibacterial activity, pharmacokinetic properties and therapeutic use." *Drugs*, no. 39, pp. 264–307, 1990.
[24] L. Gonzalez 3rd and J. P. Spencer, "Aminoglycosides: a practical review," *Am Fam Physician*, no. 15, November 1998.
[25] N. D. Goswami, C. D. Pfeiffer, J. R. Horton, K. Chiswell, A. Tasneem, and E. L. Tsalik, "The state of infectious diseases clinical trials: a systematic review of clinicaltrials.gov," *APLoS One*, October 2013.