Network Modeling Analysis in Health Informatics and Bioinformatics (2021) 10:47
https://doi.org/10.1007/s13721-021-00321-7

ORIGINAL ARTICLE

Community and topic modeling for infectious disease clinical trial recommendation

Magdalyn E. Elkin[1] · **Xingquan Zhu**[1]

Received: 12 January 2021 / Revised: 24 May 2021 / Accepted: 25 May 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2021

Abstract
Clinical trials are crucial for the advancement of treatment and knowledge within the medical community. Although the ClinicalTrials.gov initiative has resulted in a rich source of information for clinical trial research, only a handful of analytic studies have been carried out to understand this valuable data source. Analysis of this database provides insight for emerging trends of clinical research. In this study, we propose to use network analysis to understand infectious disease clinical trial research. Our goal is to understand two important issues related to the clinical trials: (1) the concentrations and characteristics of infectious disease clinical trial research, and (2) recommendation of clinical trials to a sponsor (or an investigator). The first issue helps summarize clinical trial research related to a particular disease(s), and the second issue helps match clinical trial sponsors and investigators for information recommendation. By using 4228 clinical trials as the test bed, our study investigates 4864 sponsors and 1879 research areas characterized by Medical Subject Heading (MeSH) keywords. We use a network to characterize infectious disease clinical trials, and design a new community-topic-based link prediction approach to predict sponsors' interests. Our design relies on network modeling of both clinical trial sponsors and keywords. For sponsors, we extract communities with each community consisting of sponsors with coherent interests. For keywords, we extract topics with each topic containing semantic consistent keywords. The communities and topics are combined for accurate clinical trial recommendation. This transformative study concludes that using network analysis can tremendously help the understanding of clinical trial research for effective summarization, characterization, and prediction.

Keywords Link prediction · Network community · Recommendation · Clinical trials

1 Introduction

Clinical trials carry out tests on human participants w.r.t. different interventions, including new medications or treatment, to understand and answer meaningful clinical questions (Friedman et al. 2015; Elkin and Zhu 2021). These studies are critical for discovering new treatments to diagnose, treat, and reduce the risk of disease. Understanding the concentrations and characteristics of clinical trials in specific disease areas is important for researchers and industry to be aware of emerging trends. In addition, understanding clinical trial topics can also recommend clinical trials to researchers, using shared knowledge, such as common interests, community, and topics (Hurtado et al. 2016). For example, a recommendation engine can recommend relevant clinical trials to a researchers, by using shared study topics, so the researchers can be fully aware of existing/previous studies in the field.

1.1 Clinical trial reports

The ClinicalTrials.gov database serves as a registry and results database for clinical trials all over the world. The database provides patients, researchers and public easy access to past, current and future clinical trials. In 1997, ClinicalTrials.gov was created as a registry of clinical trial information for federally and privately funded trials. The Food and Drug Administration Amendments Act (FDAAA) was created in 2007 that defines Applicable Clinical Trials (ACT) that are legally required to register their trial to the ClinicalTrials.gov database. ACTs include the following:

✉ Xingquan Zhu
xzhu3@fau.edu

Magdalyn E. Elkin
melkin2017@fau.edu

[1] Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA

Published online: 07 July 2021

clinical investigations of any U.S. Food and Drug Administration (FDA) regulated drug/biological product, certain studies of FDA-regulated medical devices, investigational studies that have one or more sites in the US, FDA investigational new drug studies and trials involving drug/biological/medical devices manufactured in U.S. (Zarin et al. 2016). While the FDAAA act specifies what clinical trials are legally required to register on the online ClinicalTrials.gov database, many trials are not legally obligated. As of October, 2020 the ClinicalTrials.gov database holds 355,127 clinical trials from over 217 countries (ClinicalTrials.gov 2020).

The ClinicalTrials.gov database is an abundant source of clinical trial studies with longest history and largest complete data (Yang and Lee 2018). Unfortunately the database is an underutilized information source for the health industry and life science research (Glass et al. 2014). Conversely, there's been an exponential increase in the amount of registered clinical trials in the US. The global clinical trial market is expected to reach 65.2 billion dollars by 2025, growing at a compound annual growth rate of over 5.5% from 2017 to 2025 (CenterWatch Staff 2017). This growth rate is expected to increase due to the growing prevalence of diseases and incidence of new diseases. Such a trend naturally raises questions on how to better analyze and utilize existing clinical reports to benefit industry, academia, and individuals (Califf et al. 2012).

1.2 Network analysis of clinical trial reports

Determining the relationship and projection of trends in clinical trials can be a daunting task considering there are large number of variables from different sub-domains. 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 (Gundogan and Kaya 2017).

For example, a previous research modeled clinical trials as a collaboration network to understand relationships between listed pharmaceutical companies, research institutes, and universities, and their mechanisms (Yang and Lee 2018). Another study created a bipartite graph from clinical trial reports from ClinicalTrials.gov to study patterns of interventions in depression trials. The authors transformed the bipartite network into a single-mode network, where intervention nodes would connect if they co-occurred in a clinical trial (Bhavnani et al. 2010). This method was able to group together similar intervention methods while quantifying trends in depression interventions.

Network analysis is commonly used for drug repurposing research. Drug or disease networks can be created using expression patterns, disease pathology, protein interactions or genetic data to find potential drugs to treat a disease of interest (Pushpakom et al. 2019). Such analysis can classify gene-disease associations with high accuracy; or identify drugs that have an effect against respiratory viral host targets (Pushpakom et al. 2019). A previous study created a disease-drug bipartite network, where a drug is connected to a disease if it's in the top three most commonly used treatments for the disease. Using an internal link based link prediction method, the authors were able to predict drugs that treated the diseases in the network (Gundogan and Kaya 2017).

In our previous study, we proposed to use bipartite network to represent clinical trial research entities and their relationships, and designed a community-based link prediction (CLP) to model sponsors as communities and predict links for information recommendation (Elkin et al. 2019). Although effective, CLP cannot make recommendation to all sponsors, because a small portion of sponsors may be assigned to invalid communities, due to their sparse connections or specialized areas not sharing by many others. As a result, sponsors in invalid communities cannot leverage information from other peers, within the same community, for recommendation.

1.3 Contributions

In this paper, we propose to use both communities and topics for clinical trial recommendation. For sponsors, we extract communities with each community consisting of sponsors with coherent interests. For keywords, we extract topics with each topic containing semantic consistent keywords. By introducing topic-based link prediction, we're able to connect sparse research areas by topics which provide a better similarity metric to compare sponsors against. The communities and topics are combined for accurate clinical trial recommendation.

The main contribution of the study is as follows.

- Infectious Disease Clinical trial Network: Our research uses a network to characterize infectious disease clinical trials and understand relationships between different factors. The network and relevant materials are published online to benefit the community.
- Community and Topic Combined Modeling: Our research proposes to simultaneously discover communities and topics to characterize sponsors and research areas in clinical trials. The combined approach delivers a solutions to connect sparse research areas or sponsors to related groups for recommendation.
- Clinical Trial Recommendation: Our research proposes to use community and topic combined link prediction

to recommend clinical trials to sponsors. The general framework can be extended to many other disease types or medical domains.

2 Data

In our study, 4228 infectious disease clinical trial reports are downloaded, in XML format, from ClinicalTrials.gov database as test bed. The downloaded reports include past, current, and future clinical trials during 1991–2023.

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 infectious disease 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. 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. Research areas are Medical Subject Headings (MeSH) Terms which roughly define the focused research topics of 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 (Huang et al. 2011). 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, and multiple research areas.

Formally, we use s to denote a sponsor 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). From our testbed, we extracted 4864 investigators (i.e. $|\mathbb{S}|$=4864) and 1878 research areas (i.e. $|\mathbb{K}|$=1878)

3 Methods

3.1 Bipartite graph for clinical trial sponsor-area relationship modeling

Clinical trials involve complex sponsors and research area relationships. A sponsor may be interested in multiple closely related (or interdisciplinary) research areas and results from one research area may be beneficial to another areas. 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 $\mathbb{G}$, ($\mathbb{V}_1 \cap \mathbb{V}_2$). In our research, sponsors represent one set of nodes and research areas represent the second set of nodes. An edge $e(s, k)$ connects a node s in sponsor node set to a node k in the research area node set ($\mathbb{E} \subset \mathbb{V}_1 \times \mathbb{V}_2$), and $\mathbb{E}$ denotes the edge set of the graph. An example bipartite network is shown in Fig. 1a. 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 Fig. 1a, deg(s_1) = 3.

If a clinical trial had multiple sponsors, edges are created from all investigators to research areas. For each edge, $e(s, k)$, a weight value $w_{s,k}$ represents the number of times an investigator is connected to a research area. To decrease the sparsity of the network, MeSH words that

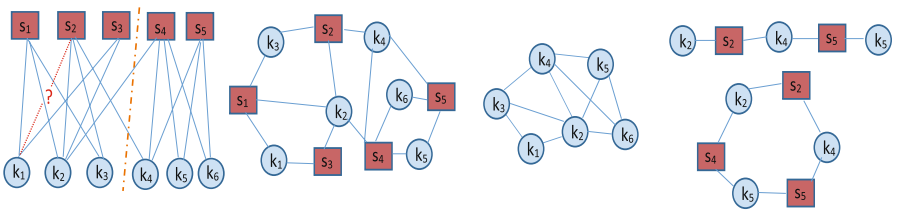


Fig. 1 A conceptual view of bipartite graph for clinical trial sponsor-area relationship modeling. **a** Shows a bipartite network where upper pink squares denote sponsors and lower blue circles indicate research areas. A blue solid line denotes an edge, indicating that a sponsor has conducted a clinical 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 a close 4-path (lower) and an open 4-path. A close 4-path in (**d**) is a circle in the one-mode network in (**c**)

contain a comma were separated into two research areas, *e.g.*, "Influenza, Human" was separated into "Influenza" and "Human".

3.2 Clinical trial network community detection

Community detection aims to find connected groups of nodes within a network. In Fig. 1a 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 (Beckett 2016). Communities are found by distinct modules that consists of a combination of two node types in a weighted bipartite network. The goal is to maximize the modularity score for a weighted bipartite network, Q_W, defined in Eq. (1) (Beckett 2016; Dormann and Strauss 2014).

$$\begin{aligned} Q_W &= \frac{1}{M}\sum_{u=1}^{\mathbb{S}}\sum_{v=1}^{\mathbb{K}}(\tilde{W}_{uv} - \tilde{E}_{uv})\delta(s_u, k_v) \\ &= \frac{1}{M}\sum_{u=1}^{\mathbb{S}}\sum_{v=1}^{\mathbb{K}}\left(\tilde{W}_{uv} - \frac{y_x z_v}{M}\right)\delta(s_u, k_v) \end{aligned} \tag{1}$$

where s and k are node types, sponsors and research areas, s_u is a sponsor node and k_v is a research area node. The Kronecker delta function $\delta(s_u, k_v)$ equals one when nodes s_u and k_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 algorithm computes modules based on two stages. 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. For a sponsor node x, its node label, s_x, is found by maximizing Eq. (2). Labels are updated until modularity score, Q_W, no longer increases (Beckett 2016).

$$\begin{aligned} s_x &= \sum_{v=1}^{\mathbb{K}}\left(\tilde{W}_{xv} - \frac{y_x z_v}{M}\right)\delta(x, k_v) \\ &= \left(\sum_{v=1}^{\mathbb{K}}\tilde{W}_{xv}\delta(x, k_v) - \sum_{v=1}^{\mathbb{K}}(\tilde{W}_{xv} - \frac{y_x z_v}{M})\delta(x, k_v)\right). \end{aligned} \tag{2}$$

In the second stage, groups of communities are merged together. Each module consists of nodes sharing the same label. Communities are merged if merging increases network modularity. This is repeated until merging more communities does not increase network modularity further (Beckett 2016). Each community, $\mathbb{C}_c$, contains a distinct subset of s and k such that $\mathbb{V} = \mathbb{C}_1 \cap \mathbb{C}_2$.

Because infectious disease clinical trials cover many diverse research areas, it is important to determine the robustness of communities. The transitivity of social networks has been widely studied (Newmann 2001; Opsahl 2011). 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 (Newmann 2001). 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 (Newmann 2001; Opsahl 2011)

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

This is frequently used in one-mode networks, an example of one-mode network is shown in Fig. 1c. 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 cannot be applied to two-mode networks, such as a bipartite network (Fig. 1a). 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 (Opsahl 2011), as shown in Fig. 1b. 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 the one where the first and last node do not connect. In Fig. 1d (upper panel) nodes k_2, s_2, k_4, s_5, k_6 are on an open 4-path. A closed 4-path (also called a 4-cycle) is a path where the first and last nodes connect. In a bipartite graph, they are connected by a 5th node. In Fig. 1d (lower panel) nodes k_2, s_2, k_4, s_5, k_5 are on a closed 4-path, closed by s_4. 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 (Opsahl 2011).

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

The second measure is 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 (Robbins and Alexander 2004).

$$RC_c = \frac{\#\text{ of closed 3-paths}}{\#\text{ of 3-paths}} \tag{5}$$

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 Fig. 6b). In this case, we consider this type of community as an invalid community.

3.3 Community-based clinical trial recommendation

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 Fig. 1a 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, such as similarity-based, supervised learning based, or collaborative filtering-based approach, have been used for link prediction (Liben-Nowell and Kleinberg 2007). In the following, we first discuss existing collaborative filtering-based link prediction, and then propose our community-based link prediction.

3.3.1 GLP: global link prediction using collaborative filtering

User-based collaborative filtering is generally performed to predict the votes of a user on a particular item by comparing the user to other users in a dataset Δ, where other users have a vote on the particular item (Breese et al. 1998). 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). Thus we are predicting $P_{s,k}$ which would indicate the weight value for sponsor s on research area k, as defined in Eq. (6). The highest value $P_{s,k}$ for k would indicate the top one predicted research area and so on.

$$P_{s,k} = \bar{v}_s + \kappa \sum_{i=1, s_i \in \Delta}^{|\Delta|} \omega(s,i)(v_{i,k} - \bar{v}_i) \tag{6}$$

In Eq. (6), Δ denotes a dataset used to determine sponsor s's sore, and $|\Delta|$ is the number of sponsors in Δ. $\omega(s,i)$ 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. (7) ($\mathcal{N}_i$ denotes the set of research area nodes connecting to sponsor s_i) (Breese et al. 1998).

$$\bar{v}_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} v_{i,j} \tag{7}$$

In summary, $P_{s,k}$ denotes sponsor node s weight on research area k. $P_{s,k}$ 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 a and b. Assume $\boldsymbol{a}$ and $\boldsymbol{b}$ are the vector representation of the sponsor of interest (a) and sponsor to compare (b) from, where $\boldsymbol{a}$ and $\boldsymbol{b}$ each denotes an m dimensional vector. The similarity between sponsors a and b is calculated as follows.

$$\omega(a,b) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{||\boldsymbol{a}|| \cdot ||\boldsymbol{b}||} = \frac{\sum_{i=1}^{m} \mathbf{a}_i \mathbf{b}_i}{\sqrt{\sum_{i=1}^{m} (\mathbf{a}_i)^2} \sqrt{\sum_{i=1}^{m} (\mathbf{b}_i)^2}} \tag{8}$$

3.3.2 CLP: community-based link prediction

In our study, we have observed that sponsor-area relationship has strong community ties, where sponsors/investigators are very likely to be interested in research areas within the same community. This is mainly because that biomedical research has a strong domain requirements, 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 clinical trials involve teams of experts from multiple domains, which essentially complicate the community structure in clinical trials.

Motivated by the above observations, we proposed a community link prediction method (CLP) to recommend links (Elkin et al. 2019), which includes three 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. The theme of the CLP is to rely on the community to recommend clinical trials areas to each sponsor. For each sponsor s, CLP uses Eq. (6) to find the sponsor's potential interest on keyword k, by using all sponsors in the same community as s, with the similarity between two sponsors calculated using Eq. (8). For each sponsor s, its vector representation $\boldsymbol{s}$ is keyword-based using Eq. (9), where $e(s,k_i)$ denotes an edge connecting sponsor s to keyword k_i, and w_{s,k_i} denotes weight of edge $e(s,k_i)$.

$$\boldsymbol{s} = \left[\mathbf{f}_1, \cdots, \mathbf{f}_i, \cdots, \mathbf{f}_m\right]; \mathbf{f}_i \leftarrow \begin{cases} w_{s,k_i}, & \text{If } e(s,k_i) \in \mathbb{E} \\ 0, & \text{Otherwise} \end{cases} \tag{9}$$

Limitation: Although effective, CLP suffers from two major limitations: (1) If a sponsor does not belong to a valid community, it cannot make recommendation to the sponsor, because there is no other sponsors in the same community to calculate similarity scores (using Eq. (8)). In our experiments, about 25% sponsors are placed in invalid communities, therefore cannot find recommendations for them; and (2) If a sponsor has a very specific focus on some rare keywords not shared by many others, CLP may not recommend accurately or fails (because of the sparsity).

To overcome the above two limitations, we propose a community-topic-based recommendation algorithm, which replies on communities and topics for recommendation. The employment of the topics ensures that keywords with low/rare occurrences are connected to others through topics, so we can make accurate recommendation to sponsors with very specific research interests, as shown in Fig. 2.

Algorithm 1 CTP: Community-Topic 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}.
$\mathbb{G}^k \leftarrow \mathbb{G}$ {Create one-mode keyword network from $\mathbb{G}$}
$\mathbb{T} \leftarrow \mathbb{G}^k$ {Find topics from $\mathbb{G}^k$}
$\mathbf{T}_{i,j} \leftarrow \mathbb{T}$ {Create sponsor-topic matrix using Eq. (11)}
repeat
  $\mathcal{Q}_W \leftarrow$ Maximizing modularity score of $\mathbb{G}$ using Eq. (1)
  for each vertex $x \in V$ **do**
    $g_x \leftarrow$ Find its modularity-based label using Eq. (2)
    $\mathcal{G} \leftarrow \mathcal{G} \cup g_x$
  end for
until Convergence
$\mathbb{C} \leftarrow$ Find communities using modularity labels $\mathcal{G}$
for each community $c \in \mathbb{C}$ **do**
  if GC_c or RC_c are valid using Eqs. (4) and (5) **then**
    $\mathbb{C}^V \leftarrow \mathbb{C}^V \cup c$ {valid community}
  else
    $\mathbb{C}^I \leftarrow \mathbb{C}^I \cup c$ {invalid community}
  end if
end for
for each sponsor $s \in \mathbb{S}$ **do**
  if s belongs to a valid community $c_x \in \mathbb{C}^V$ **then**
    $\Delta \leftarrow s_j | s_j \in c_x$
  else if s belongs to an invalid community in $\mathbb{C}^I$ **then**
    $\Delta \leftarrow s_j | s_j \in \mathbb{C}$
  end if
  for each research area k **do**
    if $e_{s,k} \notin \mathbb{E}$ {link $e(s,k)$ does not exist} **then**
      $P_{s,k} \leftarrow$ Find sponsor s's scores *w.r.t.* area k, using Eq. (6) and topic vector in Eq. (12)
    end if
  end for
end for
Rank sponsors in $\mathcal{V}$ in descending order based on $P_{s,k}$ sores.
$\mathcal{SA}_k \leftarrow$ top-k nodes on the ranked list
return $\mathcal{SA}_k$.

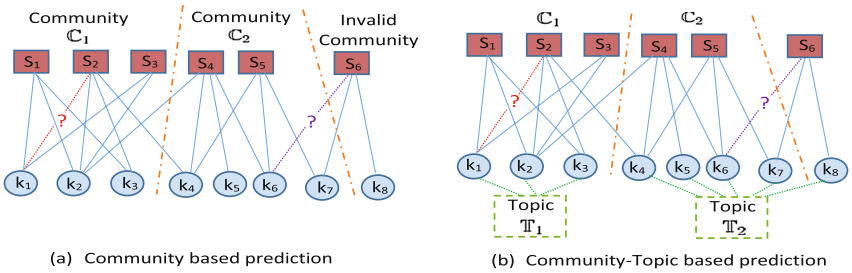


Fig. 2 Comparison between community based (Elkin et al. 2019) (**a**) vs. the proposed community-topic-based link prediction (**b**) for recommendation. Community-based approach (Elkin et al. 2019) relies on community structure for recommendation, therefore it cannot recommend link for sponsors in invalid community (e.g., the purple dashed line with a question mark from sponsor s_4 to keyword k_6). In comparison, the community-topic approach finds communities and topics from sponsors and keywords, respectively. Although sponsors s_4 is in an invalid community, the existing linkage to topic $\mathbb{T}_2$ will help recommend connection to k_6 which is within the same topic

3.4 Community-topic-based trial recommendation

3.4.1 Clinical trial topic detection

The goal of topic detection is to group keywords together as topic groups. For example, "Penicillins", "Amoxicillin", and "Ceftriaxone" could be different keywords under a topic construct, "Antibiotics". Grouping keywords as topics has two major benefits: (1) connecting different keywords at concept/semantic level; and (2) low-frequency keywords (rare search areas) can be linked to popular keywords, and tackle the sparsity challenge.

To detect topics, we use a hierarchical clustering method to combine keywords using their topological connections. To do so we create a one-mode graph, $\mathbb{G}^k$, consisting of keyword nodes only, $\mathbb{V}^k = |\mathbb{K}|$. Edges $\mathbb{E}^k$ connect two keywords, k_i and k_j, if both keywords co-occur in the same clinical trial report. Accordingly, a co-occurrence matrix $\mathbf{A}^k$ is created as the weighted adjacency matrix to create one-mode keyword graph $\mathbb{G}^k$. $\mathbf{A}^k$ is a symmetric $m \times m$ matrix, where m is the number of keywords. $\mathbf{A}^k_{i,j}$ equals to the number of clinical trials containing two keywords, k_i and k_j. $\mathbf{A}^k_{i,j} = 0$ if no clinical trial contains both keywords. An example of a one-mode keyword network is shown in Fig. 1c.

The Walktrap algorithm is applied to the keyword graph $\mathbb{G}^k$ to group keyword nodes into topic groups (Pons and Latapy 2005). WalkTrap determines structural similarities between nodes using random walks, which start on a randomly selected vertex and moves randomly in the network by following edges. The distance between two vertices is determined from the random walk, two vertices within the same subgraph region will have a small distance. Similar nodes are merged together to form communities, the merging process continues until all nodes are merged together.

To merge nodes into clusters, graph $\mathbb{G}^k$ is first separated into m clusters consisting of a single vertex. Each iteration merges two clusters into one cluster to create a new partition of the graph. This completes when all nodes are joined into one cluster. This is an agglomerative hierarchical clustering algorithm. After i iterations, there is a sequence of partitions, $P_1 \leq P_i \leq P_m$. Where P_1 is the partition of m clusters consisting of a single node and P_m is the partition of 1 cluster consisting of m nodes. Of these partitions, there is one that has the best separation of clusters, this is determined by modularity $Q(P)$ as defined in Eq. 10. Where e_C is the number of edges inside cluster C and a_C is the number of edges connected to cluster C. The partition with maximum $Q(P)$ will have the final node cluster structure (Pons and Latapy 2005).

$$Q(P) = \sum_{C \in P} e_C - a_C^2 \tag{10}$$

This method can produce cluster groups that consist of only a single node. If merging a node into a cluster decreases the partitions modularity score, $Q(P)$, it might not be desired. In our research, we wanted to group together keywords into topics to have a more generalized groupings of keywords. Thus if we had topic groups with only one node, those were ultimately merged into the cluster with the highest co-occurrence frequency. A topic cluster would consist of only a single keyword if that keyword didn't appear in any other clinical trial.

3.4.2 CTP: community-topic-based link prediction

To leverage the topics for clinical trial recommendation, we propose a community-topic-based link prediction model, which combines topics and communities for recommendation, as shown in Algorithm 1.

Topic detection: To leverage topics, we first create a topic matrix, $\mathbf{T} \in \mathbb{R}^{n \times \tau}$, with sponsors as rows and topic groups as columns. The value of $\mathbf{T}_{i,j}$, represents sponsor s_i's interest on topic $\mathbb{T}_j$, which is the total number of keywords in $\mathbb{T}_j$ having an edge to sponsor s_i.

$$\mathbf{T}_{i,j} = \sum_{k \in \mathbb{T}_j} |e(s_i, k) \in \mathbb{E}| \tag{11}$$

After the topic matrix is generated from the network, each row of the $\mathbf{T}$ is used as the the vector representation of a sponsor, as defined in Eq. (12), with Eq. (8) being used to calculate similarity between sponsors. So the similarity between two sponsors is based on the similarity on common topics, instead of based on common keywords like CLP does.

$$\boldsymbol{s}_i = \left[\mathbf{T}_{i,1}, \cdots, \mathbf{T}_{i,j}, \cdots, \mathbf{T}_{i,\tau}\right] \tag{12}$$

Community detection: Similarly to CLP, CTP also leverage the community to calculate a sponsor s's potential interest on keyword k by using all sponsors in the same community. In other words, a sponsor s is only compared to the sponsors within the same community to calculate their similarity using Eq. (8), which is calculated using vectors from the topic matrix, $\mathbf{T}$. However, due to limitations of community detection algorithms and uniqueness of network structures, not all sponsors will be placed in valid community. Therefore, CTP combines topics and communities to make recommendations for all sponsors.

Community-topic combined recommendation: Due to high sparsity and specification, invalid communities often consist of sponsors that are linked to keywords with low occurrences. Utilizing topic groups, these keywords are clustered in a method that can more accurately describe the similarity between two sponsors. From the previous study, it was determined that CLP has a high accuracy in recommendation, so if a sponsor is within a community, we still follow community-based link prediction. On the other hand, if a sponsor is within an invalid community (which consists of a single sponsor), we would use topic based global link prediction to recommend areas for the sponsor, as shown in Algorithm 1 (steps 28–39).

3.5 Time complexity

Algorithm 1 mainly includes two major components: (1) create bipartite network and find community label for each node (line 3 to 27); and (2) recommend research areas to each sponsor (line 28 to 41). Denotes $n_s = |\mathbb{S}|$ the number of sponsor nodes and $n_k = |\mathbb{K}|$ the number of research area nodes.

In order to find communities, Eq. (1) requires $\mathcal{O}(n_s \cdot n_k)$ complexity to calculate the combinations between n_s sponsors and n_k research area nodes. For each repetition, Eq. (2) needs to be calculated for all nodes $(n_s + n_k)$ in order to find their node labels. Therefore, the complexity for each repetition is $\mathcal{O}(n_s \cdot n_k \cdot (n_s + n_k))$. The process repeats until it reaches its convergence. Assume it repeats ℓ times, the complexity is $\mathcal{O}(\ell \cdot n_s \cdot n_k \cdot (n_s + n_k))$.

For recommendation, the loop from lines 28 to 39 is repeated for each sponsor and each research area, so the total complexity is $\mathcal{O}(n_s \cdot n_k)$. The sorting on line 40 is based on all sponsor and keyword pairs, and the complexity is $\mathcal{O}(n_s \cdot n_k \cdot \log(n_s \cdot n_k))$.

Because the log function $\log(n_s \cdot n_k)$ has a lower order complexity than linear function $n_s + n_k$, the complexity of the system is asymptotically bounded by $\mathcal{O}(\ell \cdot n_s \cdot n_k \cdot (n_s + n_k))$.

4 Experiments

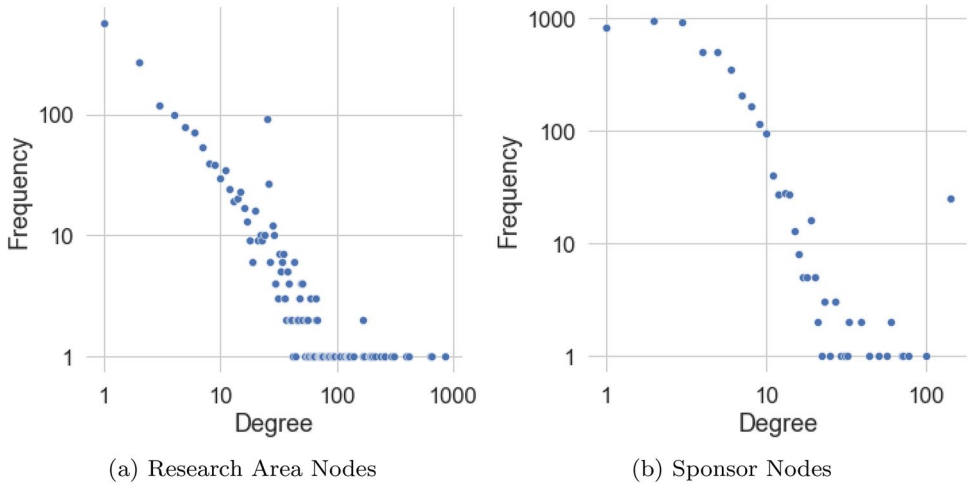
4.1 Clinical trial bipartite network characteristics

The degree distributions of Research area nodes, $\mathbb{K}$; and sponsor nodes, $\mathbb{S}$, are shown in Fig. 3a, b respectively. Both degree distributions follow a scale-free degree distribution with long-tail phenomenon. This indicates that a majority of sponsors focus on a few research areas and some research areas are studied by multiple different sponsors.

For research area nodes, the maximum $\deg(k)$ is 864. For sponsor nodes, the maximum $\deg(s)$ is 140. In total, there are 25 sponsors who all $\deg(s) = 140$. These sponsors are all connected to the same set of nodes, indicating they may have worked together on one or many clinical trials.

Table 1 lists the top 20 k nodes by degree. The top k by degree, $(\deg(k) = 864)$, is "Infection." The top 20 k nodes represent research areas in infectious disease research that receive a lot of attention from many sponsors. High on the top 20 k nodes list are "HIV Infections","Acquired Immunodeficiency Syndrome"(AIDS), "Malaria", and "Tuberculosis". These represent the "big three" infectious diseases, Malaria, tuberculosis (TB) and HIV/AIDS (Bourzac 2014). These three diseases combined accounted for 2.7 million deaths worldwide in 2018 (Prudêncio and Costa 2020). Hepatitis-related research areas ("Hepatitis","Hepatitis A", "Hepatitis C"), are also ranked high by degree. Hepatitis is responsible for 1.44 million deaths globally (Chen et al. 2015). The high ranking of these serious infectious diseases reflect serious research efforts to combat these diseases.

Research areas such as "Anti-Bacterial Agents", "Vaccines", "Antibiotics", "Antitubercular", all represent



(a) Research Area Nodes (b) Sponsor Nodes

Fig. 3 Degree distributions in log-log scale. The $x-$axis denotes node degrees, and the $y-$axis denotes the number of numbers with the specified node degrees: **a** research area nodes ($\mathbb{K}$) and **b** sponsor nodes ($\mathbb{S}$)

Table 1 Top 20 research area nodes by degree

Research area node	Degree	Research area node	Degree
Infection	864	Toxemia	212
HIV Infections	656	Hepatitis C	203
Communicable Diseases	637	Human	197
Tuberculosis	412	Influenza	193
Pneumonia	399	Respiratory Tract Infections	190
Hepatitis	309	Acquired Immunodeficiency Syndrome	172
Sepsis	295	Chronic	170
Malaria	259	Vaccines	168
Anti-Bacterial Agents	256	Antibiotics	166
Hepatitis A	235	Antitubercular	166

interventions that are ranked high as these are commonly used to treat/prevent infectious diseases. These research areas are broad and ranked high towards their likely combinations with disease research areas. Such as a sponsor may be researching vaccine development for HIV.

While the research area nodes with large degree represent those with a lot of research attention, research area nodes with a smaller degree give information on research areas that are often overlooked. The majority of k nodes have $\deg(k) < 10$, with median $\deg(k) = 3$. Often research areas with smaller degree represent more uncommon/rare infectious disease research areas. However, in some cases, it has been shown that certain infectious diseases are disproportionately neglected. The Neglected Tropical Diseases (NTDs) represent a group of infectious diseases that are commonly found in low-income developing areas of the world. These diseases affect the poorest one-sixth of the world's population and have been neglected by research attention and funding. The more recent focus on the "big three" further declined efforts towards these diseases (Feasey et al. 2010).

To compare and contrast research areas that are receive differing amounts of research efforts, Table 2 displays research area nodes of infectious diseases that are considered NTD (CDC 2020; Feasey et al. 2010). These research areas represent infectious diseases that affect 0.1 million (Trypanosomiasis) to 740 million people (Hookworm Infections) (Feasey et al. 2010). The number of years lost to disability and premature death, Disability-Adjusted Life-Years (DALYs), for NTDs is estimated at 56.6 million; compared to 84.5 million for HIV/AIDS, 46.5 million for Malaria and 34.7 million for Tuberculosis (Hotez et al. 2007). As suggested by their classification as NTDs, the

Table 2 Neglected tropical disease research areas

Research area node	Degree	Research area node	Degree
Leishmaniasis	39	Cysticercosis	7
Schistosomiasis	25	Hookworm Infections	6
Dengue	23	Onchocerciasis	5
Chagas Disease	22	Rabies	4
Leprosy	13	Severe Dengue	4
Filariasis	12	Trypanosomiasis	4
Helminthiasis	11	Echinococcosis	3
Taeniasis	7	Trachoma	3
Buruli Ulcer	7	Treponemal Infections	1

degrees of the research areas in Table 2, demonstrate the lower research efforts.

4.2 Clinical trial topic detection results

Figure 4 shows a portion of the dendrogram resulted from the clustering process with four topics (denoted by different colors). The first topic cluster, denoted by green color, represent keywords all conceptually related to facial paralysis. The keywords Facial Paralysis, Paralysis and Facial Nerve Diseases all describe facial paralysis. The first keyword, Bell Palsy, is a form of facial paralysis (de Almeida et al. 2014).

The second topic group (colored in blue) represents keywords conceptually related to neuropathic pain. Postherpetic Neuralgia is neuropathic pain due to complications caused by Herpes Zoster Oticus virus, also known as shingles (Forbes et al. 2015). Pregabalin is a treatment for postherpetic neuralgia (Derry et al. 2019). The trigeminal nerve is responsible for facial sensation; Trigeminal nerve injuries cause neuropathic facial pain (Edvinsson et al. 2020).

The third topic group, denoted by purple color, are all serious bacterial or viral infections. Tetanus, Diptheria and Pertussis, also known as whopping cough, are commonly vaccinated together with the DTaP vaccination. Recently, a new vaccine (DTaP-IPV-Hib-HepB) was approved by the FDA to prevent Diphtheria, Tetanus, Pertussis, Polio, *Haemophilus Influenzae* type B and Hepatitis B (Oliver 2020).

The final topic group, denoted by pink color, represent three serious viral diseases. Mumps, Rubella and Measles are all RNA viruses. Rubella is also known as German measles. Mumps and Measles have unique symptoms between them, but the same vaccine, measles-mumps-rubella (MMR) vaccine immunizes against all three viral diseases (White et al. 2013).

Conceptually the keywords in these topic groups represent a logical clustering. The keywords that are linked together at a lower height, such as Bell Palsy, Facial Paralysis and Paralysis, indicate these words frequently appeared in the same clinical trials. Facial Nerve diseases is a keyword that appeared with the first three keywords in only one clinical trial, thus the clustering is merged at a higher height.

After applying clustering method to keyword graph $\mathbb{G}^k$, $\tau = 169$ topic groups are derived, and the topic groups

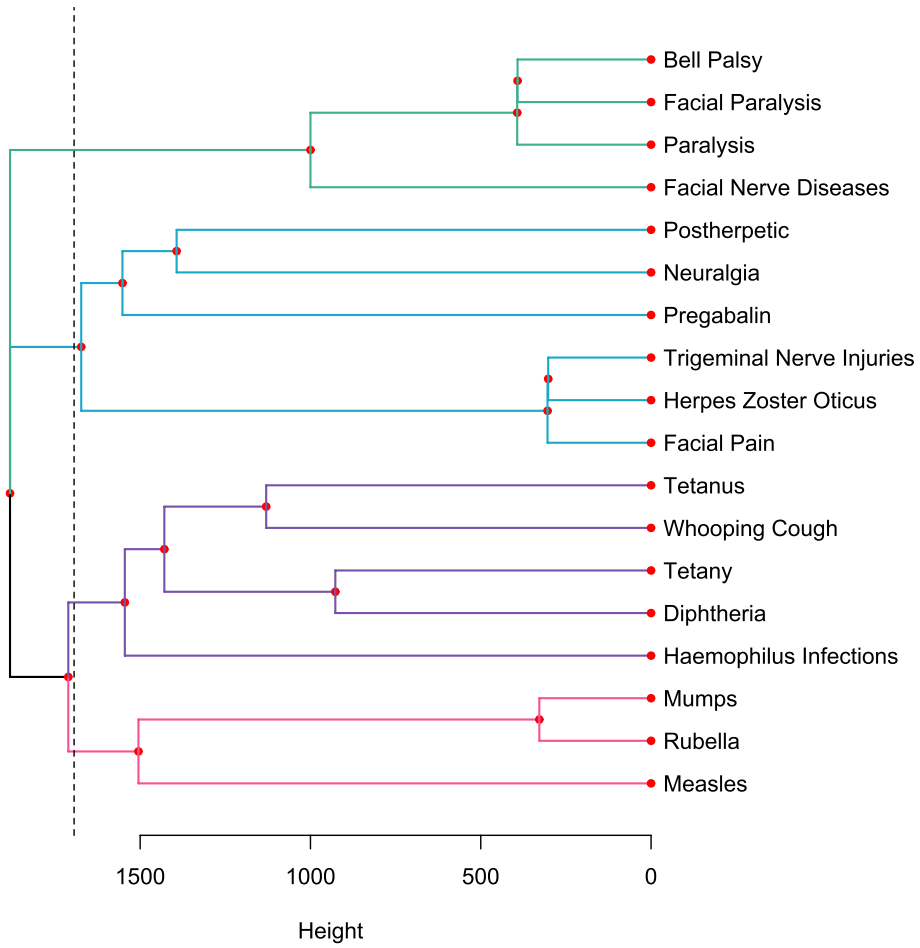


Fig. 4 A sub-dendrogram at height 2000. Four topic clusters shown. Red dots indicate when clusters were merged. The dashed line represents where the final partition separated the four clusters

have an average of 13 keywords per group. The largest topic includes 540 keywords, whereas the smallest contains 1 keyword, (there are 14 single keyword topic groups); which happens when a keyword only appears once in the set of clinical trials and there are no other keywords belonging to that clinical trial.

Figure 5 reports word clouds in two relatively large topic groups. Figure 5a are related to an oncology construct with 236 keywords such as "Lymphoma", "T-Cell", "B-Cell", etc. The topic group also contains some treatments for cancer such as Hydrocortisone. Figure 5b shows a word cloud for $\mathbb{T}_{30}$ which consists of 36 keywords. These keywords represent an HIV treatment construct.

Table 3 reports 10 selected small topic groups, the topic construct, and the respective keywords within the topic and their frequency within all clinical trials. The construct is a possible scientific construct for the keywords in the topic group, since the ultimate ground truth of the groupings of keywords is unknown and left to interpretation. Since the keywords are grouped into topics ultimately based on co-occurrence, there are some cases where an odd keyword falls within a topic group. For example, all keywords in $\mathbb{T}_{17}$ are related to the urinary system. While logically, "Stress" doesn't directly relate to the Urinary system, it can play a role in overactive bladders (Lai et al. 2015). Since the keyword "Stress" didn't appear in any other clinical trial report, it is ultimately clustered in $\mathbb{T}_{17}$.

Overall, topic detection results show that topics are useful in finding a group of keywords sharing similar/related semantic concepts. This is particularly beneficial in connecting sparse keywords to related groups, so our method can recommend trials to sponsors with high research specificity.

(a) $\mathbb{T}_{12}$ with 236 keyword (b) $\mathbb{T}_{30}$ with 36 keywords

Fig. 5 Word clouds for keywords in two separate topic groups. $\mathbb{T}_{12}$ **a** represents keywords within an oncology construct. $\mathbb{T}_{30}$ **b** represents keywords within an HIV treatment construct

Table 3 A subset of topic groups, their possible construct descriptor, and the respective keywords. The numbers beside the keywords represents the frequency of keyword, *k*, found in all clinical trials

Topic	Construct	Keywords (frequency)
$\mathbb{T}_4$	Neuropathic	Postherpetic (5); Neuralgia (7); Trigeminal Nerve Injuries (1)
		Facial Pain (1); Pregabalin (1); Herpes Zoster Oticus (1)
$\mathbb{T}_{17}$	Urinary System	Urinary Bladder (5); Overactive (3); Dyspareunia (2)
		Enuresis (1); Stress (1); Solifenacin Succinate (1)
		Urinary Incontinence (1)
$\mathbb{T}_{18}$	Hand, Foot	Mouth Diseases (3); Hand (3); Foot-and-Mouth Disease (3)
	Mouth Disease	Magnesium Sulfate (2); Foot and Mouth Disease (3)
$\mathbb{T}_{42}$	Infectious	Whopping Cough (12); Diptheria (5); Tetanus (4)
	Disease	Tetany (1); Haemophilus Infections (1)
$\mathbb{T}_{46}$	MMR	Measles (3); Mumps (1); Rubella (1)
$\mathbb{T}_{50}$	Sinus Infections	Triamcinolone (2); Triamcinolone diacetate (2)
		Frontal Sinusitis (1); Triamcinolone Acetonide (2)
		Triamcinolone hexacetonide (2)
$\mathbb{T}_{52}$	Gastrointestinal	Stomach Ulcer (4); Anorexia (2)
	System	Weight Loss (2); Duodenal Ulcer (1)
$\mathbb{T}_{55}$	Vitamin A	Vitamin A (3); Night Blindness (1)
		Retinol palmitate (3); Vitamin A Deficiency (1)
$\mathbb{T}_{61}$	Blood Clots	Mastoiditis (2); Intracranial (1); Thrombophilia (1)
		Lateral Sinus Thrombosis (1); Sinus Thrombosis (1)
$\mathbb{T}_{69}$	Facial Paralysis	Paralysis (3); Bell Palsy (3)
		Facial Paralysis (3); Facial Nerve Diseases (1)

4.3 Clinical trial community detection results

Table 4 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 4). In total, all valid communities have 3,662 sponsor nodes (s) (75.38% 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 the network. 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.

To show the structure of valid vs. invalid communities in the network, Fig. 6 demonstrates two separate communities. Figure 6a displays a valid community, $\mathbb{C}_{34}$ with 12 s nodes and 6 k nodes. Figure 6b displays an invalid community, $\mathbb{C}_{413}$ with 2 s nodes and 3 k nodes. Table 5 lists the research area nodes for the two communities.

The valid community, $\mathbb{C}_{34}$, as shown in Fig. 6a, has all closed 4-paths, thus $GC_{34} = 1$. The reinforcement coefficient is slightly lower, $RC = 0.4$, due to the four k nodes that only have connections to one other s node in the community. This indicates less localized clustering between sponsor nodes within $\mathbb{C}_{34}$.

Figure 6b displays an invalid community, $\mathbb{C}_{413}$. This community has two s nodes and three k nodes. Since one of the s nodes is only connected to one k node in the community, there is no 4-paths, thus $GC_{413} = \text{NA}$. There exists a 3-path, but it is not closed, thus $RC_{413} = 0$. Our analysis shows that a typical invalid community consists of only one or two sponsors from a single clinical trial.

Table 4 Summary of community detection results

	$\lvert\mathbb{C}\rvert$	$\lvert s\rvert$	$\lvert k\rvert$	GC_c	RC_c
Valid	139	3662	1303	.981	.054
Invalid	339	1202	575	NA	NA

Each column represents: (1) valid *vs.* invalid communities, (2) number of communities ($\lvert\mathbb{C}\rvert$), (3) number of sponsors ($\lvert s\rvert$), (4) number of research areas ($\lvert k\rvert$), (5) average Global Coefficient (GC_c), and (6) reinforcement coefficient (RC_c), respectively

Table 5 Keywords within two communities

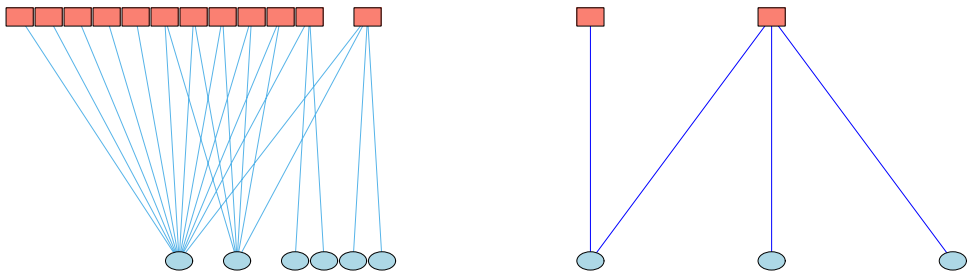
Community	Keywords
$\mathbb{C}_{34}$	Antibodies; Monoclonal; Immunological; Yellow Fever
	Blocking; Antineoplastic Agents
$\mathbb{C}_{413}$	Stomach Ulcer; Anorexia; Weight Loss

4.4 Clinical trial recommendation results

To validate the performance of the proposed clinical trial recommendation algorithm, we carry out following designs to remove a small portion of connections from the networks as benchmarks, and then compare different methods' performance in predicting 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 3b).
- $\mathbb{S}_{(6,10]}$: 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. The maximum degree of a sponsor node is 140, so up to 70 edges are removed within this node set.



(a) Valid Community $\mathbb{C}_{34}$ (b) Invalid Community $\mathbb{C}_{413}$

Fig. 6 The structure of a valid community and an invalid community: **a** valid community, $\mathbb{C}_{34}$, consists of 18 nodes ($|s| = 12, |k| = 6$); and **b** invalid community, $\mathbb{C}_{413}$, consists of of 5 nodes ($|s| = 2, |k| = 3$) The pink squares indicate sponsors and the blue circles indicate research areas

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. After creating the subnetwork with removed edges, the corresponding topic matrix is created for recommendation. If a method predicts a research area that was previously removed, 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.

In the experiments, we employ the following baseline methods for comparisons:

- **GLP and CLP:** These two methods are from our previous study (Elkin et al. 2019), which use keywords and communities for recommendation.
- **CTP** The proposed method which combines communities and topics for recommendation.
- **CTP** $_t$: A variant of the proposed CTP method, which removes the community detection module and only uses topics for recommendation.

The purpose of using CTP$_t$ is to carry out ablation study and remove communities to study CTP's performance. Alternatively, CLP only uses communities and does not use topics. Therefore, by comparing CLP *vs.* CTP$_t$, we can understand whether topics are playing more important role than communities for recommendation, or vice versa. It is worth noting that CLP relies on communities for recommendation, so it only works on sponsors within valid communities. GLP, CTP, and CTP$_t$, on the other hand, work for all sponsors.

Because some methods only work for valid communities, we carry out experiments by comparing their performance on Valid Community Network (which only consists of sponsors within valid community) and all network, respectively.

4.4.1 Link prediction on valid community network

To compare the performance of GLP, CLP, CTP, and CTP$_t$, three benchmark node sets are created on a subnetwork consisting of nodes only in valid communities. In each benchmark node set, 10 sponsor nodes are selected and half their links are removed with each method being used to predict links for recommendation. This repeats 20 times, and the mean accuracy for the link prediction methods is reported in Figs. 7 and 8.

Figure 7a, shows the performance with respect to top 3 accuracy. This benchmark node set, $\mathbb{S}_{[2,6]}$, is the only case that CLP marginally outperforms CTP. GLP has the worst performance, and the addition of topic-based link prediction (CTP$_t$) increases accuracy.

Figure 7b shows the performance with respect to top 5 accuracy. As the number of removed edges increases from

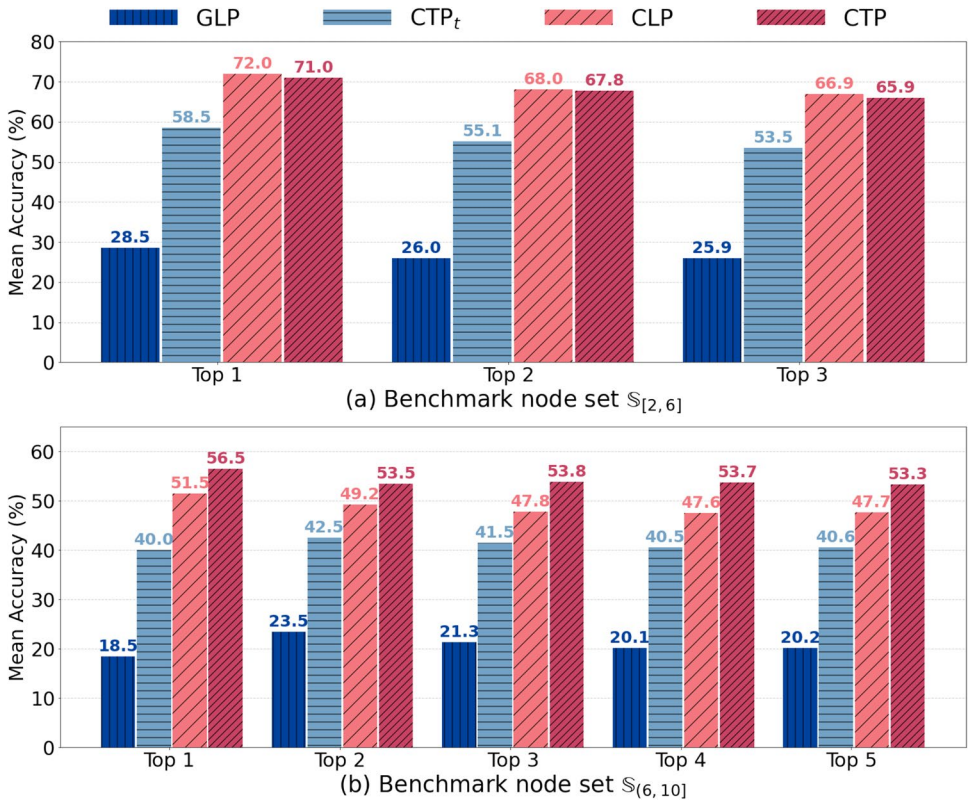


Fig. 7 Link prediction accuracy comparison on valid community networks, **a** using benchmark node set $\mathbb{S}_{[2,6]}$, and **b** using benchmark node set $\mathbb{S}_{(6,10]}$. The x−axis denotes the top-k prediction, and the y−axis denotes the link prediction accuracy

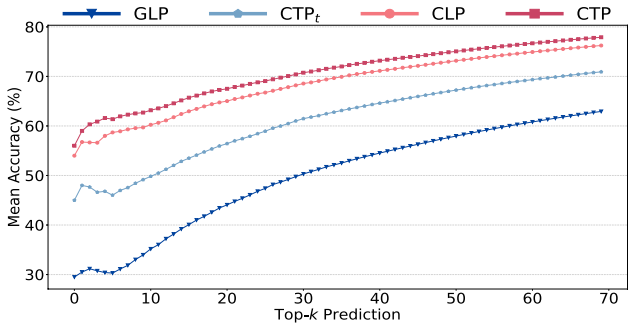


Fig. 8 Link prediction accuracy on the valid community network using benchmark node set $\mathbb{S}_{(10,\infty)}$. The x−axis denotes the top-k prediction, and the y−axis denotes the accuracy

maximum 3 to maximum 5, overall accuracy decreases. CTP now outperforms CLP.

Figure 8 shows the performance with benchmark node set $\mathbb{S}_{(10,\infty)}$. The maximum degree of a sponosor node is 140, thus sponsors in this node set had up to 70 edges removed. Generally, as the number of edges removed increases, the accuracy increases for all methods. CLP and CTP have similar performance, CTP achieves higher accuracy, and CTP$_t$ achieves much better performance than GLP.

The results can be summarized into two major findings: (1) the addition of Topic-based link prediction significantly increases recommendation accuracy; and (2) the usage of community-based link prediction increases accuracy.

4.4.2 Link prediction on the whole network

The whole network consists of all nodes, regardless if they are within a valid community or not. Due to the findings that topic based and community-based increases link prediction accuracy, we wanted to validate our CTP method on the whole network. To do so we create three benchmark node sets as previously described, in each node set, 20 sponsor nodes are selected and half their edges are randomly removed. Then GLP, CTP, and CTP$_t$ are used to predict links. This is repeated 20 times. The average accuracy of link prediction for the three benchmark node sets is reported in Figs. 9 and 10.

Figure 9a shows the accuracy with respect to top-3 prediction with benchmark node set $\mathbb{S}_{[2,6]}$. The addition of topic-based link prediction, CTP$_t$, shows a major improvement compared to GLP. The addition of community and topic-based link prediction, CTP, shows the highest performance.

Figure 9b demonstrates the accuracy with respect to top-5 prediction with benchmark node set $\mathbb{S}_{(6,10]}$. As with the valid community network, this benchmark node set has slightly decreased performance for all methods. GLP shows the lowest performance. CTP shows an increased advantage over CTP$_t$.

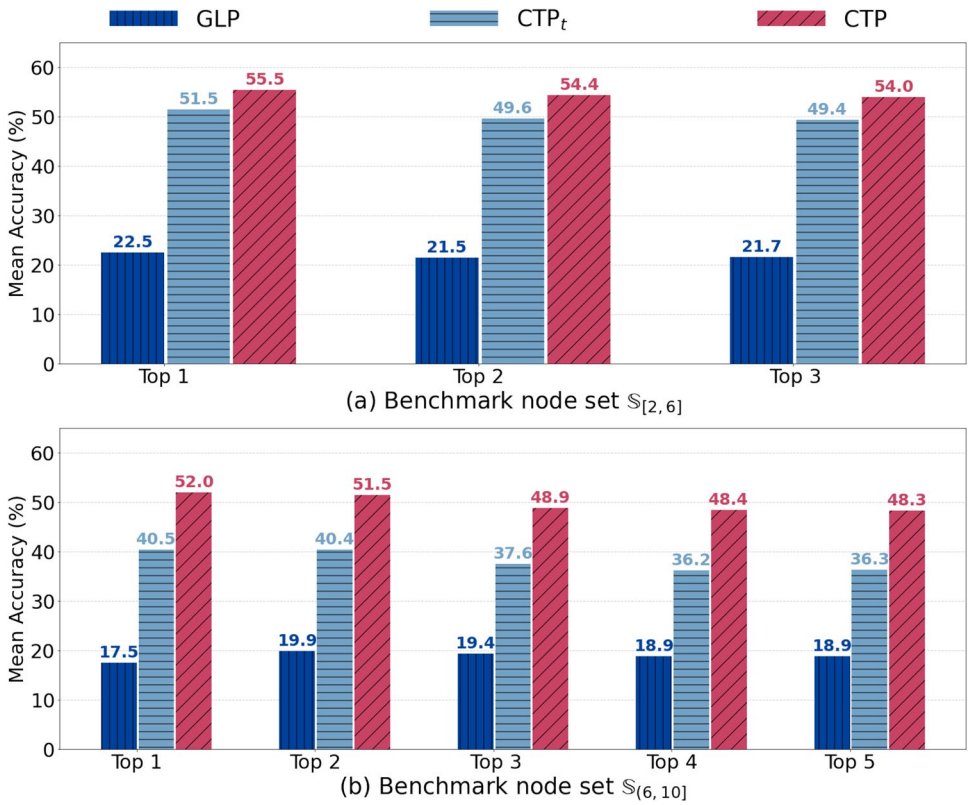


Fig. 9 Link prediction accuracy comparison on valid community networks, **a** using benchmark node set $\mathbb{S}_{[2,6]}$, and **b** using benchmark node set $\mathbb{S}_{(6,10]}$. The x−axis denotes the top-k prediction, and the y−axis denotes the link prediction accuracy

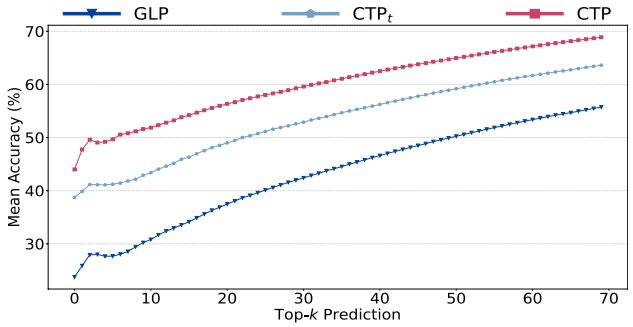


Fig. 10 Link prediction accuracy on the whole network using benchmark node set $\mathbb{S}_{(10,\infty)}$. The x–axis denotes the top-k prediction, and the y–axis denotes the accuracy

Figure 10 demonstrates the accuracy with benchmark node set $\mathbb{S}_{(10,\infty)}$ with respect to Top 1 to Top 70 prediction. This follows the trend in the valid community network with increasing accuracy as the number of edges is removed. Again CTP has the highest performance and GLP has the lowest performance.

Overall the results show that the addition of topic-based link prediction increases accuracy. A topic-based similarity metric provides a better basis for similarity comparison between two sponsors increasing the accuracy of collaborative user filtering.

5 Discussions

This study aims to characterize clinical trials using network analysis of sponsors and research areas. By modeling infectious disease clinical trials as a bipartite network between sponsors and research areas, we can differentiate infectious disease research areas receiving a lot *vs.* a little attention. The degree of a research area directly measures the number of sponsors studying the infectious disease. While it is expected for more uncommon infectious diseases to receive a smaller degree of research efforts, some infectious disease research areas do receive disproportionate research efforts. Our results show that NTDs infectious disease research areas have considerably smaller degree compared to the "big three" (HIV/AIDS, Malaria, Tuberculosis) and Hepatitis. This demonstrates their classification as "Neglected". Similar analysis on research areas with a small degree can identify commonly overlooked infectious diseases.

Our previous research method demonstrated the high predictive power of using community-based link prediction (Elkin et al. 2019). While CLP does have an increase in performance, it still can't be used effectively towards the whole network. CLP requires that all nodes exist within a valid community. This study expands on our previous method by introducing topic-based modeling. By finding topics based on keywords and summarizing the number of keywords in each topic per sponsor, this metric provides a better basis for similarity comparison between sponsors increasing the accuracy of collaborative user filtering. Finding topics effectively groups together some keywords that are not as common within the network. If a sponsor only has connections to uncommon keywords, grouping them together can accurately represent the similarity between two sponsors. This is demonstrated in the increased performance of CTP$_t$ compared to GLP within both the valid community network and the whole network. The introduction of topic-based similarity is a more reliable similarity metric.

As shown with invalid community, C_{413}, the network connections are sparse, indicating that using community structure for link prediction wouldn't be accurate for invalid communities. Invalid communities often only consist of 1–2 s nodes. With sparse connections, community-based link prediction may be unreliable if there are only a small number of sponsors in the community. The keywords for C_{413}, as listed in Table 5, are also found within topic $\mathbb{T}_{52}$. The groupings of keywords within a topic can provide more useful information regarding sponsors connected to these research areas than relying on community structure only.

The performance of CTP within the valid community network is greater than CTP within the whole network. Since the performance of GLP is similar between valid community network and whole network, we can conclude that the network size isn't the determining factor. This demonstrates the high predictive power of link prediction within a valid community network. However, this exclusion of nodes is not always feasible, especially if the sponsor of interest belongs to an invalid community. The superior performance of CTP *vs.* CTP$_t$ demonstrates the power of using community information, if a node does belong to a valid community, using community-based link prediction will increase accuracy. Ultimately, CTP has increased performance because it utilizes both community and topic information.

For both whole network and valid community network, performance on benchmark node set $\mathbb{S}_{(6,10]}$ is slightly reduced than benchmark node set $\mathbb{S}_{[2,6]}$. The majority of sponsor nodes ($|s| = 4056$) fall within benchmark node set $\mathbb{S}_{[2,6]}$. These sponsors may represent those who only have specialized or localized research interests. As deg(s) increases for nodes in benchmark node set $\mathbb{S}_{(6,10]}$, the research areas become broader and link prediction accuracy is slightly reduced for all methods. In benchmark node set, $\mathbb{S}_{(10,\infty)}$, accuracy increases as the number of links predicted increases. As the deg(s) increases for a sponsor node, the likelihood increases that the node belongs to a highly localized dense community. For example, the 25 s nodes that all have deg(s) = 140 all belong to a large dense community with 46 sponsors and 170 research areas. The

dense connections effectively provide more information for each sponsor node and increase the link prediction accuracy, resulting in a gradual increase in accuracy for all methods. The difference between GLP and CTP is greatest at lower Top-k predictions. This demonstrates the ability of CTP to rely more necessary information regarding a sponsor node, which is necessary when the connections are less dense.

Overall these results suggest that link prediction has increased benefits from researchers in localized/specialized areas and researchers with large degrees (i.e. many research areas shared by many other researchers). Meanwhile, link prediction shows a decline for researchers with a broader set of interests while maintaining a lower degree.

In our research, the topics are based on the graph $\mathbb{G}^k$, instead of using node content. This indicates the original dataset itself has high importance with regards to finding topics. For example, if more clinical trials contained the keyword "Stress", that would affect the keyword's placement in a topic group. Within the dataset used for this study, "Stress" only was found in one clinical trial, which determined it's placement into a Urinary System topic group construct, as shown in Table 3. Using more information to enrich the networks can essentially improve the topic discovery, and result in more accurate clinical trial recommendation.

6 Conclusions

In this study, we proposed to study relationships between investigators/sponsors and research areas in infectious disease clinical trials extracted from ClinicalTrials.gov. which is a valuable, but under utilized, data source. We used bipartite graph to create infectious disease networks between sponsors and research areas, and studied characteristics of the networks. The analysis of research area degree demonstrates the research efforts given to separate infectious diseases. Our research shows that clinical trial research follows unique scale-free network characteristics: (1) researchers are highly specialized where many of them primarily work on specific research areas, although a handful a researchers indeed work on many areas; (2) a small number of research areas are very commonly studied by many researchers, yet many research areas are studied by a small number of researchers. Overall, infectious disease research for the "big three" and Hepatitis receive large research efforts/attention from sponsors, whereas infectious disease research for NTDs receive a smaller amount of sponsor attention.

For accurate clinical trial recommendation, we proposed to reduce sparsity in the data, by extracting communities to group sponsors and using topics to model research areas. Combining communities and topics, we formed a link prediction task to recommend research areas for sponsors. Experiments and validations confirmed that, compared to the previous research, the proposed method is much more accurate in recommending links for infectious disease clinical trial research. The proposed method provides an accurate and reliable method for recommending clinical trial research areas to a sponsor.

Future research can emphasize on integrating additional relationships, such as drug keywords, into the network analysis, or extending the proposed framework to other clinical trial areas, such as heart disease.

Acknowledgements This research is sponsored by the US National Science Foundation through Grants IIS-2027339, IIS-1763452, CNS-1828181.

References

Beckett SJ (2016) Improved community detection in weighted bipartite networks. R Soc Open Sci 3:140536

Bhavnani SK, Carini S, Ross J, Sim I (2010) Network analysis of clinical trials on depression: implications for comparative effectiveness research. AMIA Annu Symp Proc 2010:51–55

Bourzac K (2014) Infectious disease: beating the big three. Nature 507(7490):S4–S7. http://www.nature.com/articles/507S4a

Breese JS, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., Madison, Wisconsin, pp 43–52

CDC (2020) Diseases - Neglected Tropical Diseases. Available: https://www.cdc.gov/globalhealth/ntd/diseases/index.html

Califf RM (2012) Characteristics of clinical trials registered in ClinicalTrials.gov, 2007–2010. JAMA 307:1838

Chen D-S, Locarnini S, Wallace J (2015) From the big three to the big four. Lancet Infect Dis 15:626–627

ClinicalTrials.gov (2020) Trends, charts, and maps. https://clinicaltrials.gov/ct2/resources/trends#RegisteredStudiesOverTime

de Almeida JR, Guyatt GH, Sud S, Dorion J, Hill MD, Kolber MR, Lea J, Reg SL, Somogyi BK, Westerberg BD, White C, Chen JM, Bell Palsy Working Group, Canadian Society of Otolaryngology – Head and Neck Surgery and Canadian Neurological Sciences Federation (2014) Management of Bell palsy: clinical practice guideline. CMAJ 186:917–922

Derry S, Bell RF, Straube S, Wiffen PJ, Aldington D, Moore RA (2019) Pregabalin for neuropathic pain in adults. Cochrane Database Syst Rev 1:01

Dormann C, Strauss R (2014) A method for detecting modules in quantitative bipartite networks. Methods Ecol Evol 5:90–98

Edvinsson JCA, Viganò A, Alekseeva A, Alieva E, Arruda R, De Luca C, D'Ettore N, Frattale I, Kurnukhina M, Macerola N, Malenkova E, Maiorova M, Novikova A, Řehulka P, Rapaccini V, Roshchina O, Vanderschueren G, Zvaune L, Andreou AP, Haanes KA, On behalf of the European Headache Federation School of Advanced Studies (EHF-SAS) (2020) The fifth cranial nerve in headaches. J Headache Pain 21:65

Elkin ME, Zhu X (2021) Predictive modeling of clinical trial terminations using feature engineering and embedding learning. Sci Rep 11:3346

Elkin ME, Andrews WA, Zhu X (2019) Network analysis and recommendation for infectious disease clinical trial research. In: Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics. ACM, Niagara Falls, NY, pp 347–356

Feasey N, Wansbrough-Jones M, Mabey DCW, Solomon AW (2010) Neglected tropical diseases. Br Med Bull 93(1):179–200. https://academic.oup.com/bmb/article-lookup/doi/10.1093/bmb/ldp046

Forbes H, Thomas S, Smeeth L, Clayton T, Farmer R, Bhaskaran K, Langan S (2015) A systematic review and meta-analysis of risk factors for postherpetic neuralgia. Pain 157:07

Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB (2015) Fundamentals of Clinical Trials, 5th edn. Springer, New York

Glass HE, Glass LM, DiFrancesco JJ (2014) Clinicaltrials.gov:an underutilized source of research data about the design and conduct of commercial clinical trials. Therap Innovt Regul Sci 49(2):218–224

Gundogan E, Kaya B (2017) A link prediction approach for drug recommendation in disease-drug bipartite network. In: 2017 International artificial intelligence and data processing symposium (IDAP). IEEE, Malatya, pp 1–4

Hotez PJ, Molyneux DH, Fenwick A, Kumaresan J, Sachs SE, Sachs JD, Savioli L (2007) Control of neglected tropical diseases. N Engl J Med 357:1018–1027

Huang M, Névéol A, Lu Z (2011) Recommending MeSH terms for annotating biomedical articles. J Am Med Inform Assoc 18:660–667

Hurtado LJ, Agarwal A, Zhu X (2016) Topic discovery and future trend forecasting for texts. J Big Data 3:7

Lai H, Gardner V, Vetter J, Andriole GL (2015) Correlation between psychological stress levels and the severity of overactive bladder symptoms. BMC Urol 15:14

Liben-Nowell D, Kleinberg J (2003) The link prediction problem for social networks. In: Proceedings of the twelfth international conference on Information and knowledge management - CIKM '03. ACM Press, New Orleans, LA, pp 556–559

Newman MEJ (2001) Scientific collaboration networks. I. Network construction and fundamental results. Phys Rev E 64:016131

Oliver SE (2020) Licensure of a diphtheria and tetanus toxoids and acellular pertussis, inactivated poliovirus, haemophilus influenzae type b conjugate, and hepatitis B vaccine, and guidance for use in infants. MMWR Morb Mortal Wkly Rep 69:136–139

Opsahl T (2011) Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. arXiv:1006.0887 [physics.soc-ph]

Pons P, Latapy M (2005) Computing communities in large networks using random walks. In: Yolum pInar, Güngör T, Gürgen F, Özturan C (eds) Computer and information sciences - ISCIS 2005. Springer, Berlin, pp 284–293

Prudêncio M, Costa JC (2020) Research funding after COVID-19. Nat Microbiol 5:986–986

Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, McNamee C, Norris A, Sanseau P, Cavalla D, Pirmohamed M (2019) "Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discovery 18(1):41–58. http://www.nature.com/articles/nrd.2018.168

CenterWatch Staff (2017) Report: Global clinical trials market is expected to reach 65.2B by 2025. Available: https://www.centerwatch.com/news-online/2017/09/25/report-global-clinical-trials-market-expected-reach-65-2b-2025/

Robbins G, Alexander M (2004) Small worlds among interlocking directors: network structure and distance in bipartite graphs. Comput Math Org Theory 10:69–94

White SJ, Boldt KL, Holditch SJ, Poland GA, Jacobson RM (2013) Measles, mumps, and rubella. Clin Obstet Gynecol 55:550–559–922, 06

Yang H, Lee HJ (2018) Long-term collaboration network based on clinicaltrials.gov database in the pharmaceutical industry. MDPI Sustain, pp 1–14

Zarin DA, Tse T, Williams RJ, Carr S (2016) Trial reporting in clinicaltrials.gov–the final rule. N Engl J Med 375:1998–2004

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.