Heidelberg University
Institute of Computational Linguistics

Analyse von Netzwerken zwischen Pharma-Firmen sowie von klinischen Studien auf die Frage, ob befreundete Firmen noch ihre Produkte gegeneinander testen

Analyzing a pharma company network: are friends still competing?

Bachelor's thesis
supervised by Prof. Dr. Stefan Riezler and Prof. Dr. Gerhard Reinelt

by
Michael Haas
Immatriculation number: 2775430
haas@cl.uni-heidelberg.de
24th October 2011

Thanks to:
My parents, for making everything possible.
Dr. Katharina Zweig, for having good answers and even better questions.

EIDESSTATTLICHE ERKLÄRUNG

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht sind und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Leimen, den 24. Oktober, 2011 Michael Haas

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, that I have explicitly marked all material which has been quoted either literally or by content from the used sources and that this work has not been submitted, either in part or whole, for a degree at this or any other University.

Leimen, 24th of October, 2011 Michael Haas

1. German Abstract

Wir überprüfen die Hypothese, ob befreundete Medikamentenhersteller ihre Medikamente weniger häufig vergleichen, um beispielsweise ihre Freunde im Wettbewerb nicht zu benachteiligen. Die Eigenschaft der "Freundschaft" zwischen Herstellern wird gekennzeichnet über gemeinsame Insider wie Anteilseigner oder Vorstandsvorsitzende.

Das der Analyse zugrunde liegende Netzwerk wird aus verschiedenen Quellen erstellt, deren Daten zusammengeführt werden. Die Daten über Firmen und ihre Insider stammen von der amerikanischen Handelsaufsichtsbehörde, der SEC. Hier werden auf einer Website Liste von Entitäten - Firmen und Personen - geführt, die jemals Insider einer Firma waren. Hier nutzen wir die Daten aller Firmen, die gemäß Kategorisierung nach Standard Industrial Code unter "pharmaceutical preparations" (Code 2834) eingeordnet sind. Die Daten über medizinische Versuche stammen von der Website `http://www.clinicaltrials.gov/`, wo die National Institutes Of Health der USA Meta-Daten zu klinischen Studien bereitstellen. Hier nutzen wir alle Interventionsstudien aus dem Zeitraum 1.1.2006 bis 31.12.2010. Die Daten, um die Interventionen, also die verwendeten Medikamente, aus den Studien deren Herstellern zuzuordnen, stammen aus DrugBank (Knox et al., 2011), einem Projekt der University of Alberta.

1.1. Verknüpfung der Daten

Die Daten der drei Quellen müssen verknüpft werden. Es gilt, die Liste der Firmen der Handelsaufsichtsbehörde mit der Liste der Hersteller aus DrugBank abzugleichen. Hierbei stellt sich das Problem, dass die gleiche Firma mit leicht unterschiedlichem Namen referenziert sein kann. Eine selbst geschriebene Lösung zur Verknüpfung der Einträge liefert einen F-Score von $\sim 0,8$.

Ein ähnliches Problem stellt sich bei der Verknüpfung von Interventionen der klinischen Studien mit den Namen der Medikamente aus DrugBank. Neben unterschiedlichen Namen für Medikamente können Interventionen zusätzlich aus mehreren Medikamenten bestehen, so dass einzelne Medikamente extrahiert und Rauschen wie Angaben zur Dosierung gefiltert werden müssen. Auch hier liefert eine eigene Lösung einen F-Score von ~ 0.8.

Das so erstellte Netzwerk umfasst 22162 Kanten und 16679 Knoten, die sich wie folgt verteilen: 561 Medikamente, 5152 medizinische Versuche, 7288 Vergleichsgruppen der Versuche, 164 Hersteller und 3514 Insider.

1.2. Analyse der Daten

Das Netzwerk wird zur Analyse in zwei Perspektiven, sogenannte *One-mode projections*, auf die Menge der Firmen zerlegt. Die Insider-Perspektive zeigt, wie die Insider ein Freundschaftsnetzwerk zwischen Firmen aufspannen. Die Studien-Perspektive zeigt, welche Firmen ihre Medikamente in klinischen Studien vergleichen.

Zwecks Überprüfung der Hypothese betrachten wir alle befreundeten Firmen, also alle Firmen-Paare mit Abstand 1 in der Insider-Perspektive, und betrachten die Häufigkeit

des Ereignisses, dass ein solches Paar seine Medikamente vergleicht. Im Vergleich mit zufälligen Paaren wird offensichtlich, dass befreundete Firmen eine geringere Wahrscheinlichkeit haben, ihre Medikamente in wenigstens einer Studie zu vergleichen zu vergleichen ($p < 0,7\%$).

Die Perspektiven auf die Menge der Firmen können, bedingt durch die naive Methode der Projektion, insignifikante Verknüpfungen enthalten. Aus diesem Grund testen wir für beide Perspektiven getrennt die Signifikanz von Verbindungen zwischen Firmenpaaren gegen ein Null-Mode, basierend auf der Arbeit von Zweig (2010), und erzeugen gefilterte Netzwerke bei verschiedenen Signifikanzniveaus.

Der oben beschriebene Effekt wird auf den meisten gefilterten Perspektiven weiterhin offensichtlich; bei manchen Signifikanzniveaus ist der Effekt weiterhin sichtbar, aber nicht statistisch signifikant.

1.3. Zusammenfassung

Unsere Beobachtung, dass befreundete Firmen weniger häufig ihre Medikamente in wenigstens einer Studie vergleichen als zufällige Paare von Herstellern, unterstützt unsere Hypothese. Andererseits lässt die Beobachtung auch den Schluss zu, dass Insider - Investoren - absichtlich Firmen aus verschiedenen Sparten aussuchen, um beispielsweise vor Umsatzeinbrüchen einer Sparte gefeit zu sein. Da in der Regel nur Medikamente für ähnliche Indikationen verglichen werden, erklärte dies die niedrigere Frequenz.

Zudem gilt es zu beachten, dass die Zusammenfassung und Auswertung der Daten kein fehlerfreier oder verlustfreier Vorgang ist. Nur circa 15% der medizinischen Studien verbleiben nach Filterung mit dem Kriterium nur Studien zu behalten, in denen mindestens zwei Medikamente verglichen werden. Die Verknüpfung der Daten bei einem F-Score von $0,8$ bietet Raum sowohl für falsche als auch für fehlende Zuordnungen.

Dennoch sehen wir auf diesem Netzwerk interessante Zusammenhänge zwischen Medikamentenhersteller, ihren Insidern und medizinischen Versuchen, die nicht durch Zufall allein entstehen. Tatsächlich bleiben die Effekte für die Mehrzahl der gefilterten Perspektiven sichtbar.

Part I.
Introduction

2. Problem and Hypothesis

It is well-known in the medical community that pharmaceutical companies influence the drug research process to increase sales revenue. Favourable results in clinical trials increase the market share of a company, and financial conflicts of interest put the researcher's objectivity and their scientific integrity at risk. A meta analysis conducted by Bekelman et al. (2003) finds:

> [..] evidence suggests that the financial ties that intertwine industry, investigators, and academic institutions can influence the research process.

While this meta analysis is mainly concerned with pharmaceutical research, it is also common for pharmaceutical companies to influence physicians directly, as they will be choosing which drug is prescribed to the patient. Brennan et al. (2006) note that even small gifts and pharmaceutical samples tip the physician's opinion towards the company in question. They further note that 90% of the $21 billion marketing budget of the industry is directed towards physicians. Most saliently, they write that

> [the company's] ultimate fiduciary responsibility is to their shareholders who expect reasonable returns on their investments.

An older paper by Kessler et al. (1994) describes the phenomenon of "me too" drugs, where several companies compete for market shares in the same therapeutic class. In these cases, a company releases a drug similar to already existing drugs, targeting the same population. Even though the new drug may not offer a clear advantage over existing drugs, the manufacturer may still decide to bring it to the market. To this end, several tactics are described by the authors. One are "seeding trials", where a large number of office-based medical practitioners are recruited to prescribe the new drug to their patients. This tactic is used to introduce the drug to physicians while the scientific usefulness of these types of studies - unblinded, no control group - often is questionable.

Given how pharma companies influence the research process to obtain a larger market share, we want to find out if 'friendly' companies, e.g. manufacturers closely linked by common shareholders and directors, influence drug selection in clinical trials. A drug showing better performance in clinical trials is likely to have a higher market share, thus providing more revenue to the manufacturer of the winning drug. We hypothesize that friendly pharmaceutical companies avoid damaging each other's market share and revenue stream by comparing their products.

3. Structure

Starting from publicly available data, we will build a model of the relations between companies, insiders and drug tests. We will employ network analytic approaches on this model to test our hypothesis.

This article is structured as follows: after giving some definitions in Section 4, we will describe the data sources in detail in Section 5. The process of assembling the data sources into the final model, in particular the solution we employed to link records from different sources, is presented in Section 6. Our approach to test our hypothesis is based on extracting two different perspectives on the set of companies and described in Section 7. The final analysis of these perspectives is provided in Section 7.2.

As an additional step, we test the significance of relationships in our network against a null model and again verify our hypothesis against the resulting filtered graph in Section 8.

In the conclusion, we discuss our experiences with the open data efforts used in this projects and recommend some improvements which would have made our work much easier (Section 9). We present ideas for future work as well as another interesting pattern we observed in the network (Section 10.3), which warrants further investigation. Finally, we close with our results in Section 11.

Part II. Methods

4. Definitions

We start by establishing some definitions we will use in this work. The following basic network definitions mostly stem from a paper by Zweig (2010).

A *graph* consists of a set of nodes V and a set of edges $E \subseteq V \times V$. The number of nodes is denoted by $n = |V|$, the number of edges is $m = |E|$. We will also use the term *network* to refer to graphs (Newman, 2003). A *bipartite graph* is a graph whose node set V can be divided into two disjoint sets V_0 and V_1, with no edges connecting two nodes from the same set V_i: all edges (v, w) must be between nodes $v \in V_0$ and $w \in V_1$. The *co-occurrence* $cooc(v, w)$ for two nodes v, w from the same set V_i is defined as number of nodes z in the other set $V_{j \neq i}$ so that (v, z) and $(w, z) \in E$. The *degree* $deg(v)$ is the number of neighbors of a node v, i.e. the number of edges where v is participating. The *degree distribution* of a node set $W \subseteq V$ is the distribution of degrees over nodes $w \in W$. The *degree sequences* L and R of the graph are defined as the sequence of the degrees in V_0i and V_1, assuming some fixed order on the nodes.

When a bipartite graph $B = (V_0 \cup V_1, E)$ is turned into a graph $G(V_i, E')$ on one of the two node sets V_i, the resulting graph G is called a *one-mode projection.* A *clique* is a set of nodes where each node is connected to each other by an edge (Luce and Perry, 1949). A *network motif* is a "recurring, significant patterns of inter-connections"(Milo et al., 2002). The *clustering coefficient* C_v for a single node v with a degree k_v is defined as: $C_v = \frac{2t_v}{k_v \cdot (k_v - 1)}$ (Watts and Strogatz, 1998; Zahoranszky et al., 2009). t_v denotes the number of triangles in which node i participates; that is, the number of neighbors of v which are directly connected to each other. $k_v \cdot (k_v - 1)$ denotes the maximum number of triangles around node i; i.e. the maximum number of neighbors that could be connected to each other. The average clustering coefficient over the whole graph is then given by $C = \frac{1}{n} \sum_{v \in G} c_v$. As Watts and Strogatz (1998) put it, the clustering coefficient measures the "cliquishness' of the network.

We close with some informal definitions concerning the realm of clinical trials. A *clinical trial* is a "prospective study comparing the effect and value of intervention(s) against a control in human beings" (Friedman et al., 2010).

We will use the terms *study*, or *clinical study,* synonymously with *trial.* In a trial, different *interventions* consisting of e.g. drugs or medical devices are tested and compared

Figure 1: Data flow chart to illustrate the dependencies of the data source

in different *arms* or *treatment groups*. Typically, a trial "contains a control group against which the intervention group is compared"(Friedman et al., 2010).

5. Data sources

We employ different data sources to build our network, which we will describe below. A basic data flow chart to illustrate the dependencies between the data sources is presented in 5. To build a basic network representation of drug trials, we choose to use the ClinicalTrials.gov database. ClinicalTrials.gov, a service provided by the U.S. National Institute of Health, currently lists 115,179 trials with locations in 177 countries[1]. We found that this database does not provide all information we need to test our hypothesis; namely, it is missing the manufacturer of the drugs used in the trials. For this purpose, we used the DrugBank database (Knox et al., 2011) which provides a list of manufacturers for drugs along with a wide array of other information. Finally, we need to learn which companies are connected by insiders. This information is obtained from the website of the U.S. Securities and Exchange Commission. A detailed explanation of the data sources follows in this section while efforts to link the data records from different sources are described in 6.2.

We also describe a fourth data source here, secondary in nature: the DBPedia project (Auer et al., 2007). We do not use data provided by the DBPedia project to add new entities to the network; rather, we evaluated its use in the context of linking records between ClinicalTrials.gov and DrugBank.

During our research on data sources, we noticed that the Linked Data Community also provides their own versions of the data in RDF[2] format. RDF, short for Resource Description Framework, is a standard popular in the Semantic Web Community which lends itself easily to network analysis. We will present the results of our investigation in these representations and describe why we found them unfit for our purpose in Section 5.5.

[1] `http://clinicaltrials.gov/`, accessed 22.10.2006
[2] `http://www.w3.org/RDF/`

5.1. ClinicalTrials.gov

We need know which drugs are tested against each other. Based on such drug pairings, we can later look up which companies compare their drugs in clinical trials.

The U.S. National Institute of Health provides a list of clinical trials at `http://www.clinicaltrials.gov/`. The site provides the ability to list clinical trials based on various search criteria. The search results can be downloaded and provide meta data about the clinical studies in XML format.

For our purposes, we retrieve only interventional studies. Observational studies are not of interest to us as there is no direct comparison between different treatments. The data retrieved consists of all interventional studies from 01/01/2006 to 12/31/2010 (75733 studies), retrieved on 15.02.2011.

The following data is extracted from each XML document:

- unique trial ID
- download date
- trial sponsors
  - lead sponsor, collaborators and their respective agency_class, e.g. whether the sponsor class is industry or federal
- interventions, i.e. the treatments used in the study
  - intervention type, e.g drug, device, radiation
  - intervention name, e.g. the name of the drug
  - synonyms: other names for the intervention (other_name field)
  - description
  - arm group label: study arm in which this intervention was tested
- start date of the study (optional)
- end date of the study (optional)

We collect synonyms for any given intervention into a global list. If a study contributes an additional synonym for an intervention, then this intervention is added to the global list of synonyms for this intervention.

Only a small portion of studies are considered for building the network. Studies are filtered based on the following criteria:

- a study must have more than one arm. If only one arm is present, then no comparison between drugs is taking place or the comparison is not easily obvious from the meta data as it was assembled incorrectly.
- at least two arms must each have one intervention of type "drug". This avoids trials where drugs are tested against, for example, medical devices.

After this filtering step, 11942 studies or about 15% are left.

5.2. DrugBank

Unfortunately, the meta data provided by ClinicalTrials.gov does not list the manufacturers of the drugs tested in the trials. To build our network, we need to associate each drug with its manufacturers. For this purpose, we use the DrugBank database (Finucane and Boult, 2004).

The DrugBank database is "a richly annotated database of drug and drug target information" (Finucane and Boult, 2004). Among many properties of the drugs, such as their structure, function and action, it also contains a list of manufacturers and packagers for each drug. The website at `http://www.drugbank.ca/` provides both a web interface to the database as well as a complete download of the database in XML format. To access the XML data, we loaded the DrugBank 3.0 meta data into BaseX 6.5.1[3], a database for XML documents which supports the XML Query Language (XQuery)[4].

The XML schema can be found at `http://drugbank.ca/docs/drugbank.xsd`, the complete database is located at `http://drugbank.ca/system/downloads/current/drugbank.xml.zip`.

An abbreviated example for the XML follows:

```
<drug type="biotech" created="2005-06-13 07:24:05 -0600" version="3.0" updated="
    2010-11-25 15:36:58 -0700">
<drugbank-id>DB00001</drugbank-id>
<name>Lepirudin</name>
<description>Lepirudin is identical to natural hirudin[..]</description>
<cas-number>120993-53-5</cas-number>
<brands>
<brand>Refludan</brand>
</brands>
<mixtures/>
<packagers>
<packager>
<name>Bayer Healthcare</name>
<url>http://www.bayerhealthcare.com</url>
</packager>
</packagers>
<manufacturers>
<manufacturer generic="false">Bayer healthcare pharmaceuticals inc</manufacturer>
</manufacturers>
</drug>
```

It should be noted that looking up the correct manufacturer set for a given drug from a clinical trial is not trivial, as drug names need to be normalized first (see Section 3 for details).

An example query written in the XQuery language for drug Lepirudin follows:

```
declare default element namespace "http://drugbank.ca";
for $x in doc("drugbank")/drugs/drug
where ($x/name/upper-case(text())=upper-case("Lepirudin"))
or ($x/synonyms/synonym/upper-case(text())=upper-case("Lepirudin"))
or ($x/brands/brand/upper-case(text())=upper-case("Lepirudin"))
or ($x/mixtures/mixture/name/upper-case(text())=upper-case("Lepirudin"))
return $x/drugbank-id/text()
```

[3] BSD-licensed, available at `http://basex.org/`

[4] `http://www.w3.org/TR/xquery/`

Here, we perform a case-insensitive search of the *name* field along with the fields for *synonym*, *brand* and *mixture.*

5.3. Insider Data: U.S. Security and Exchange Commission

Based on the sources for clinical trial meta data and manufacturer information, we can find out which manufacturers compare their drugs in clinical trials. To test our hypothesis, however, we still need to know which companies are closely linked by insiders. To this end, we query an online database provided by the U.S. Security and Exchange Commission (SEC)[5].

The SEC oversees and regulates the U.S. securities market[6]:

> All investors, whether large institutions or private individuals, should have access to certain basic facts about an investment [..]. To achieve this, the SEC requires public companies to disclose meaningful financial and other information to the public.

The SEC also provides a system for Electronic Data Gathering, Analysis, and Retrieval (EDGAR) where companies submit their filings. EDGAR can be used by the public to retrieve facts about companies as well as a company's filings. In EDGAR, both companies and individuals are identified by the Central Index Key (CIK)[7]. Additionally, companies are categorized using a Standard Industrial Classification code (SIC)[8]. For example, searching for the SIC 2834 will list all companies categorized as *pharmaceutical preparations.*

Of all company filings submitted to the SEC, forms 3, 4 and 5 are relevant to our goals. These forms are used for ownership reports. Details from these forms are used to build the ownership report in EDGAR[9].

- Form 3 is used when stocks are given out and bought for the first time
- Form 4 is used when ownership of equity changes
- Form 5 is used in cases of deferred reporting

In any case, the insider is responsible for filing the report with the SEC. Companies or individuals that qualify as insiders are[10]:

- company officers
- company directors
- any owners of more than ten percent of a class of the company's equity securities

[5] `http://sec.gov/`
[6] `http://sec.gov/about/whatwedo.shtml`, retrieved on 01.09.2011
[7] `http://sec.gov/edgar/quickedgar.htm` accessed 01.09.2011
[8] `http://sec.gov/info/edgar/siccodes.htm`, retrieved 01.09.2011
[9] `http://sec.gov/answers/form345.htm`, retrieved 01.09.2011
[10] see Securities Exchange Act of 1934, Section 16, a1) for details

Insider definition The SEC unfortunately does not explain how the insider lists in the ownership reports are created. While it is obvious that insider status is regulated by law, it is not clear how entities are extracted from forms submitted to companies. Two inquiries sent to the SEC did not yield any results. Consider the insider list for "ACADIA PHARMACEUTICALS INC" (CIK 0001070494), which lists, among others, 12 different insiders as "10 percent owner" (as of 22.09.2011). This obviously does not add up and leads us to believe that insider lists are not necessarily updated if an entity gives up ownership.

Concrete Implementation A list of all pharmaceutical companies is extracted using the Standard Industrial Code for *Pharmaceutical Preparations* (2834). Since EDGAR provides no dedicated Application Programming Interface for such requests, the EDGAR web interface is parsed with a program written in Python 2.7 using the BeautifulSoup HTML parser[11]. Central Index Keys for the companies are extracted by navigating the DOM tree. EDGAR does not display all search results on a single page, so the python script automatically retrieves search results on additional pages.

Data on insiders is retrieved directly from EDGAR. Hence, we do not process the company filings ourselves, but rely on the processing in EDGAR.

In this work, an "insider" of a company is any entity listed under the "Get insider transactions for this issuer" link in the EDGAR web interface. There, a table is presented with a list of owners and the corresponding filing number, the transaction date and the type of owner, which will typically be a variation of either "director", "officer" or "10 percent owner".

To retrieve all insiders for a company, the Central Identification Keys are used to construct URLs pointing to the correct records. For example, the correct URL to retrieve all insiders for the equity issuer *ABBOTT BIOTHERAPEUTICS CORP* with CIK 0001441848 would be `http://www.sec.gov/cgi-bin/own-disp?action=getissuer&CIK=0001441848`. Again, information from these tables is extracted using a Python script and BeautifulSoup since no API is available.

5.4. DBPedia

As we will describe in more detail in Section 6.2, we link records from different data sources together to build the network. For the mapping from ClinicalTrials.gov interventions to the corresponding manufacturers, we attempted to leverage the DBPedia data set (Auer et al., 2007). A description of the source follows below; an evaluation of the mapping process using this option can be found in section 6.2.1. It is important to note that this data source does not contribute entities to the network. It merely aids in linking records across sources.

Wikipedia articles for drugs usually contain an *infobox* with a link to the corresponding DrugBank entry. These infoboxes are used by the DBPedia project to create a machine-readable representation of some of the information in Wikipedia in Resource Description

[11] `http://www.crummy.com/software/BeautifulSoup/`

Framework (RDF) format.

The basic building block of the RDF format is the URI[12], which represents entities as well as the relation between entities. Three URIs can be combined in a triple similar to the Subject-Verb-Object structure of natural language[13].

The RDF data is available via SPARQL[14] endpoints, which can be queried over the Internet using the SPARQL query language. Since infoboxes are usually filled in by hand by Wikipedia editors, the data quality is high. As such, one would expect to get a high quality mapping for drug names to the corresponding DrugBank ID. A typical triple set of interest in DBPedia for drug Azacitidine[15] in Notation3[16] follows:

```
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dbpedia-owl: <http://dbpedia.org/ontology/> .
dbpedia:Azacitidine rdfs:label "Azacitidine"@en;
dbpedia:Azacitidine rdf:type dbpedia-owl:Drug;
dbpedia:Azacitidine dbpedia-owl:drugbank "APRD00809"@en
dbpedia:Ladakamycin dbpedia-owl:wikiPageRedirects dbpedia:Azacitidine
```

This example first defines some aliases to increase readability of the triples. The last four lines contain the actual RDF assertions. The first assertion assigns a name to the entity dbpedia:Azacitidine, the second statement integrates dbpedia:Azacitidine into the DBPedia ontology. The third assertion describes the DrugBank ID, the last line asserts that the Wikipedia page for Ladakymacin redirects to Azacitidine, which usually is an indication that these concepts are closely related.

A sample SPARQL query to retrieve the DrugBank ID for Azacitidine looks as follows:

```
SELECT ?id WHERE {
        ?concept rdfs:label "Azacitidine"@en .
        ?concept dbpedia-owl:drugbank ?id }
```

Implementation Notes For communication with the SPARQL endpoint, the SPARQLWrapper library version 1.4.2 was used[17]. The DBPedia project provides a public endpoint at `http://dbpedia.org/sparql`; however, performance and reliability issues prompted us to set up our own copy of the triple store. Due to previous experience with the Jena Semantic Web Framework [18], we choose to use the TDB software (version 0.8.9)[19] as data store and the Joseki SPARQL server[20] (version 3.4.3) to access the data store content.

The following files from DBPedia release 3.6 were loaded into the triple store:

[12] Uniform Resource Identifier

[13] Of course, the full RDF specificiation is much more powerful, but this simple description is sufficient for our purposes

[14] SPARQL Protocol And RDF Query Language, `http://www.w3.org/TR/rdf-sparql-query/`

[15] `http://dbpedia.org/data/Azacitidine.n3`

[16] Also known as 'n3' format, `http://www.w3.org/TeamSubmission/n3/`

[17] `http://sparql-wrapper.sourceforge.net/`

[18] `http://jena.sourceforge.net/index.html`

[19] `http://openjena.org/TDB/`

[20] `http://www.joseki.org/`

- instance_types_en.nt: contains *rdfs:type* attributes
- labels_en.nt: *rdfs:label* attributes
- mappingbased_properties_en.nt: attributes extracted from infoboxes
- redirects_en.nt: contains *dbpedia-owl:wikiPageRedirects* attributes
- specific_mappingbased_properties_en.nt: contains more specific attributes extracted from infoboxes

The mappingbased_properties_en.nt file was prepared with the *riot* tool which is shipped as part of ARQ ("a SPARQL Processor for Jena", version 2.8.7)[21], before it could be loaded into the triple store, due to syntax errors.

The process for the DrugBank ID lookup for a drug x in DBPedia is as follows:

- Look up instance (Wikipedia page) with label "x" or "X"
- if instance exists, check if it is a redirect. Follow redirects up to depth 2.
- look up DrugBank ID for instance

Given the DrugBank ID, drug manufacturers are then looked up in DrugBank itself as described in Section 5.2.

5.5. Third-party representations of the data

During our investigation of the available data sources, we noticed that the Linked Data community provides representations of the ClinicalTrials.gov and DrugBank data in RDF format. We will not describe our findings on these third-party representations and why we found them unsuitable for our purposes.

The DrugBank data set is available in RDF format at `http://www4.wiwiss.fu-berlin.de/drugbank/`. However, the origin of this data set is unclear: there is no information on the conversion process from DrugBank data to RDF. There is also no versioning information. At the time of investigation, it seemed obvious that the data provided by the SPARQL endpoint did not reflect the DrugBank 3.0 release. For example, the data set lists Dexamethasone (DB01234) as having been updated in 2008, while the entry on the DrugBank website had been updated in 2011.

The data set itself is very interesting, as it also links to other resources such as DailyMed[22]. DailyMed is a project of the National Library of Medicine and provides "high quality information about marketed drugs"[23]. This interlinked data is of potential interest for informed network generation, but due to uncertainty about data quality, this data source was dismissed.

[21] `http://openjena.org/ARQ/`
[22] `http://www4.wiwiss.fu-berlin.de/dailymed/`
[23] `http://dailymed.nlm.nih.gov/dailymed/about.cfm`, retrieved 08.09.2011

Of special interested is also the RDF version of the ClinicalTrials.gov data set, which is available at `http://www.linkedct.org/`. The DrugBank RDF data set we just described is interlinked with the RDF version of ClinicalTrials.gov. In theory, this means that it is easily possible to list all trials for a given drug, which should simplify network generation greatly. However, there is no mention of how the named entity resolution problem was handled: are multi-drug interventions considered at all? If yes, how are drugs extracted and how well does this method perform? Working without this information is hardly possible as the quality of the generated network will be largely unknown.

As an additional problem, the conversion process from XML to RDF does not handle the arm_group_label foreign key properly. In proper ClinicalTrial data, it is easy to see which intervention belongs to which study arm, thus it is easy to find out which drugs are being compared. In linkedct.org data, the "arm group label" field on an intervention does not link back to the RDF entity for the arm. Instead, it provides the label as text, resulting in cumbersome queries.

As a more practical problem when working with the DrugBank RDF data, we found that the generated links to the linkedct.org database cannot be resolved. We cannot retrieve information about these entities, such as the trials in which they are participating. Thus, we are unable to look up which drugs are compared using this data source. Additionally, the DrugBank RDF data set does not contain manufacturer or packager information.

To sum up, we came to the conclusion that the third-party representations described above are not suitable for our purpose. The data is either old or of unknown quality, in particular with respect to the problem of resolving entity mentions across sources in order to link records. We believe that evaluating the methods used to build the network, as shown in Section 6, is necessary to provide a solid foundation for our analysis.

6. Building the Network

So far, we have described where to retrieve the data necessary to test our hypothesis. We need to link this data to form a network. First, we must extract drug comparisons from the data. We then must link the drug names to their manufacturers via information provided by DrugBank. The following step is to link the drug manufacturers with the insider lists provided by EDGAR.

6.1. Determining which drugs are compared

Naively, we assumed that different <intervention> tags in the ClinicalTrial.gov data each stand for a drug or a combination of drugs that is compared with the other interventions. However, this assumption proved to be false. Studies such as NCT00382590 contain three intervention fields, but only two arms. Here, participants in one arm are given two drugs. This also means that one intervention tag per drug was used. However, as indicated above, this convention does not always hold. Very often, an intervention already contains the combination of drugs given to participants in one group, so there is an 1:1 relationship between interventions and arms in some study meta data. We decided

to restricted our extraction to interventions which have an arm group label assigned, as this is a sure way of knowing that the interventions - and thus the drugs contained therein - are being compared.

6.2. Linking the records

We have extracted data from various sources:

- drugs which are compared in trials from ClinicalTrials.gov
- drug manufacturer data from DrugBank.ca
- insider data from the SEC's EDGAR system

To build the final network, these data need to be linked. Drug names from ClinicalTrials.gov needs to be linked to drug names in DrugBank.ca. Manufacturer names from DrugBank.ca need to be linked to manufacturer names from EDGAR. This problem is also known as Named Entity Resolution in the Natural Language Processing community.

6.2.1. Resolving drug name mentions between ClinicalTrials.gov and DrugBank

The meta data presented by ClinicalTrials.gov does not list single drugs. Instead, it lists interventions. An intervention can be composed of multiple drugs, or even a combination of drugs and medical devices. Information on dosage and administration may also be present in the intervention description. The result of our matching process is a mapping from ClinicalTrials.gov mentions to DrugBank mentions.

So before actually matching drug mentions, we need to find out which drugs are actually mentioned in the ClinicalTrials.gov data. The following is a list of interventions seen in the meta data:

- Etanercept, methotrexate and depomedrone
- albumin interferon alfa-2b
- SYN117 (nepicastat)
- Bevacizumab (Avastin), Taxotere (Docetaxel), Vinorelbine Tartrate (Navelbine)
- exenatide twice daily (BID)
- pemetrexed, cisplatin and erlotinib before surgery then erlotinib is given to patients after surgery for 2 years
- levofloxacin at 250, 500, 750, and 1000 mg doses
- Placebo for ABT-869
- Study drugs: Metformin and fenofibrate

Traditional pattern-matching techniques were employed to resolve this problem.

A development and test set was created by sampling 500 interventions uniformly at random from the set of eligible trials as per Section 5.1, further limiting the set to trials from 01.01.2008 to 31.12.2010. A gold standard was created by manually mapping the 250 intervention tags of the test set to zero or more DrugBank entries.

From the development set, possible delimiters are extracted. The final list of delimiters is found in Appendix A.

Please note that the patterns for the conjunctions contain spaces to avoid wrongly splitting drug names in the middle, such as 'Avandia' for the 'and' pattern.

Noise patterns were extracted which are completely removed from the tokens. The list of noise patterns is also found in Appendix A.

The algorithm for tokenization is as follows:

```
def split(intervention):
tokens = set()
# initialize token set with intervention
tokens.add(intervention)
for d in delimiters:
        temp = set();
        for t in tokens:
                # split current token around delimiter
                # then strip surrounding white space
                ts = map(lambda s: s.strip(), t.split(d))
                temp.update(ts);
        tokens = temp;
return tokens
```

The filtering simply works by matching all noise patterns against every token and replacing matches with the empty string. In case a whole token is replaced with the empty string, it is deleted from the list of tokens.

The look-up process The Look-up process uses two parameters: the source which is queried and the tokenizer.

Evaluation Two sources are available: direct DrugBank look-up using a local copy of the DrugBank XML data set and the XBase XQUERY server as well as look-up via the DBPedia data set. Additionally, combining the sources is possible in two modes: union and intersection. For example, the union between the results of the DBPedia source and the direct DrugBank look-up might increase recall. On the other hand, the intersection between these result sets might improve precision. Two tokenizers are available: one using the delimiters as described in appendix A, and a 'space' tokenizer which adds the space character (ASCII 0x20) to the list of delimiters.

A gold standard was created to find out which combination performs best. We manually created a mapping from intervention strings to DrugBank IDs. The data source for this mapping mostly was DrugBank. Occasionally, a web search engine revealed an additional synonym which lead to the correct DrugBank entry. For some drugs, no DrugBank

Source	Combination	Tokenizer	Precision	Recall	F-Score
DBPedia + DrugBank	Intersection	Default	0.9888	0.4251	0.5946
DBPedia	-	Default	0.9785	0.4396	0.6067
DrugBank	-	Default	0.9533	0.4928	0.6497
DBPedia + DrugBank	Union	Default	0.9459	0.5072	0.6604
DBPedia + DrugBank	Intersection	Space	0.9728	0.6908	0.8079
DBPedia	-	Space	0.9733	0.7053	0.8179
DrugBank	-	Space	0.9000	0.7826	0.8372
DBPedia + DrugBank	Union	Space	0.9016	0.7971	0.8462

Table 1: Results DrugBank ID look-up evaluation, sorted by F-Score in ascending order

entry could be found. The gold standard was created by building a list of all interventions of type 'drug' in the ClinicalTrial data set and then sampling 250 interventions uniformly at random from that list.

Development of the tokenization and look-up methods was done by manual inspection of the complete ClinicalTrial data set and continuously refined. No development on tokenization and look-up methods took place using direct feedback from the evaluation based on the gold standard, except for minor tweaks. The gold standard was then used to find out which of the available parameters performed best.

Given 4 possibilities for the data source (DBPedia; DrugBank; DBPedia AND DrugBank, DBPedia OR DrugBank) and two choices for the tokenizer, there are 8 combinations to be evaluated. The results are listed in table 6.2.1.

Originally, we intended to employ the intervention synonyms found in the ClinicalTrial.gov data set to improve the linking process. Unfortunately, the full data set exhibited a peculiarity not found in the evaluation data set which made this option unfeasible[24]. For reasons of data purity, we opted against modifying the full data set and instead left out the synonyms here. We provide the full evaluation results in appendix B.

It is obvious that adding the space character to our list of drug delimiters improves performance considerably as all combinations employing the space tokenizer perform better than their counterparts using the default tokenizer. As expected, the DBPedia source always provides better precision than the direct DrugBank look-up, while the latter has an advantage in recall. Overall, direct DrugBank look-up performs better than DBPedia look-up, considering the F-Score.

Thus, best results for a single source are achieved with the DrugBank source and the space tokenizer. Using the union of both data sources provides slightly better perform-

[24]For 'placebo' interventions, the other_name field in the clinical trial meta data often contains a description of the substitution taking place, describing which drug is replaced with a placebo. Since we maintain a global list of intervention synonyms where the other_name content for a given intervention is collected from all studies, this creates a lot of false positives for other_name fields like "Ventolin placebo and Spiriva placebo" (study NCT00981851). The result is that studies employing a 'placebo' control group will erroneously have many drugs linked to that group due to the way the tokenizer works.

DrugBank	EDGAR	CIK
Leiner health products inc	LEINER HEALTH PRODUCTS INC	0001043055
Apothecon inc div bristol myers squibb	BRISTOL MYERS SQUIBB CO	0000014272
Bristol myers squibb pharma co	BRISTOL MYERS SQUIBB CO	0000014272
Warner chilcott inc	WARNER CHILCOTT PLC	0001042459
Warner chilcott inc	WARNER CHILCOTT PLC	0001113445
Warner chilcott inc	Warner Chilcott CORP	0001319893
Warner chilcott inc	Warner Chilcott Holdings CO III, LTD	0001319896
Warner chilcott inc	Warner Chilcott Intermediate (Luxembourg) S.a.r.l.	0001331426

Table 2: Example mapping between DrugBank and EDGAR mentions

ance. The intersection between both sources provides very high precision while recall suffers compared to other source choices.

Given these results, we can conclude that the space tokenizer provides best performance when paired with the union of the result sets for the DBPedia and DrugBank sources. By choosing different sources, precision can be improved while penalizing recall (and vice-versa) with near-identical F-Score.

For practical purposes, we choose to use DrugBank only as data source to speed up data look-up, as the minuscule difference in F-Score (~ 0.009) does not justify the overhead of setting up and querying two data sources.

To sum up, we will use the DrugBank source and the space tokenizer to link entity mentions from ClinicalTrials.gov interventions to drug names in DrugBank. With a precision of 0.9 and a recall of 0.7826, we expect the resulting links to be of sufficient quality to serve as a base for our network analysis efforts.

6.2.2. Resolving company name mentions between DrugBank and EDGAR

We face a similar Named Entity Resolution problem when matching company names from DrugBank and EDGAR. Again, different naming conventions make it hard to link mentions to entities. Thus, we need to develop a strategy to resolve mentions and evaluate the quality of this solution. Table 2 lists some example data.

It is immediately obvious that many entries in EDGAR are written in capital letters. As the second example shows, the DrugBank manufacturer entries can contain additional information on company structure. Apothecon is a subsidiary of Bristol-Myers Squibb. As EDGAR only lists Bristol-Myers Squibb and not Apothecon, a simple string comparison will not produce a match. Again we need to rely on proper tokenization to retrieve the relevant substrings. The third example again is concerned with Bristol-Myers Squibb. Here, the additional token "pharma" is included in the DrugBank entry.

The last five examples show possible matches for DrugBank mention "Warner chilcott

inc". The upper-cased counter-part "WARNER CHILCOTT PLC" exists twice with different CIKs.

Additionally, the possible matches show that different entities in the legal sense might not be considered different in our sense. While Warner Chilcott Intermediate in Luxembourg is clearly a different legal entity than the Warner Chilcott PLC in Ireland, it is hard to tell which of these legal entities is the manufacturer listed in DrugBank. Since the legal entities are most likely tightly linked, a distinction is not necessary and perhaps detrimental to our cause.

Similar duplicates can be found in the DrugBank data set, as in example 2 and 3, where two different strings link to the same EDGAR CIK.

Tokenization Again, the tokenizer is based on string patterns collected from working with the data. The pattern list can be found in Appendix C.

The tokenizer supports different modes.

- Simply split company name along any occurrence of a pattern from the pattern list (TOKENIZER_SPLIT)
- Compile patterns into regular expressions and split the company name where the expressions match (TOKENIZER_SPLITREGEX)
- Compile delimiters into regular expressions and remove substrings matching these expressions (TOKENIZER_FILTERREGEX)
- Remove any occurence of a pattern from the company name string (TOKENIZER_FILTERDELIM)
- Split company name along blanks (ASCII 0x20) and filter out tokens matching various regular expressions compiled from pattern list (TOKENIZER_SPLITBLANK_REGEX)
- Split company name along blanks and filter out any tokens matching patterns in pattern list (TOKENIZER_SPLITBLANK_DELIM)

A single pattern 'delim' is compiled into four regular expressions for the REGEX modesl:

1. "^delim "
2. " delim$"
3. " delim "
4. "^delim$"

The behaviour of the python regular expression matcher in 'search' mode is to scan the complete input string until it can find a substring matching the regular expression. This is not desired behaviour as a short delimiter like "co" might wrongly match inside a word. Thus, we introduce anchoring to the regular expressions. For the compiled expression to match, the delimiter must either be surrounded by two spaces (case 3), start at the beginning of the string and have a trailing space (case 1), end at the end of a string and have a leading space (case 2), or simply comprise the whole string.

Token matching A candidate string from each source is tokenized and the token sets are handed to the matcher. The matcher considers two candidate strings to refer to the same entity if a single token is shared between token sets (MATCHER_EXACT). A second mode MATCHER_SEARCH is provided where one token must be a substring of another token from the other company name. As a third mode (MATCHER_EDIT), the matcher matches two tokens if their Levensthein distance[25] is 1 or lower. The general algorithm is as follows, with the behaviour of the *check_token_equivalence* method determined by the matcher mode and the behaviour of the *tokenize* method determined by the tokenizer mode, as described above.

```
def matches(company_name1, company_name2):
  tokens1 = tokenize(company_name1)
  tokens2 = tokenize(company_name2)
  for t1 in tokens1:
    for t2 in tokens2:
      if check_token_equivalence(t1, t2):
        return True
  return False
```

The general algorithm for named entity resolution is as follows:

```
def resolve(edgar_companies, drugbank_companies):
  resolved = []
  for ec in edgar_companies:
    for dc in drugbank_companies:
      if matches(ec,dc):
        # remember matches
        resolved.append((ec,dc))
  return resolved
```

Evaluation results are listed in table 3. With an F-score of 0.8081, the best method is the combination of the exact matcher (MATCHER_EXACT) and the tokenizer that works by splitting the entity mentions along regular expressions compiled from the pattern strings (TOKENIZER_SPLITREGEX).

6.3. Final Network

The final network is directed to allow for easy navigation along with node attributes. Trials are connected via directed edges to their arms which in turn have directed edges to drugs tested in these arms. The drugs link to their manufacturers using direct edges. On the other side of the network, insiders link to manufacturers in which they participate using directed edges. In this model, the only node types having no incoming edges are insider nodes and trial nodes. See figure 2 for a schematic view of the data model.

[25] nltk.metrics.distance.edit_distance function from the Python Natural Language Toolkit, `http://www.nltk.org/`

Tokenizer	Matcher	F-Score
TOKENIZER_SPLIT	MATCHER_SEARCH	0.0452
TOKENIZER_SPLIT	MATCHER_EDIT	0.0766
TOKENIZER_SPLIT	MATCHER_EXACT	0.0819
TOKENIZER_FILTERDELIM	MATCHER_EXACT	0.1724
TOKENIZER_FILTERREGEX	MATCHER_EXACT	0.2034
TOKENIZER_FILTERREGEX	MATCHER_EDIT	0.2034
TOKENIZER_FILTERDELIM	MATCHER_EDIT	0.2333
TOKENIZER_SPLITREGEX	MATCHER_SEARCH	0.3274
TOKENIZER_FILTERREGEX	MATCHER_SEARCH	0.3438
TOKENIZER_SPLITBLANK_DELIM	MATCHER_SEARCH	0.3915
TOKENIZER_SPLITBLANK_REGEX	MATCHER_SEARCH	0.3932
TOKENIZER_FILTERDELIM	MATCHER_SEARCH	0.4359
TOKENIZER_SPLITREGEX	MATCHER_EDIT	0.7207
TOKENIZER_SPLITBLANK_DELIM	MATCHER_EDIT	0.7339
TOKENIZER_SPLITBLANK_REGEX	MATCHER_EDIT	0.7767
TOKENIZER_SPLITBLANK_DELIM	MATCHER_EXACT	0.8041
TOKENIZER_SPLITBLANK_REGEX	MATCHER_EXACT	0.8041
TOKENIZER_SPLITREGEX	MATCHER_EXACT	0.8081

Table 3: Named Entity Resolution for company names between EDGAR and DrugBank

The final network is stored as GraphML file[26]. Nodes are annotated with *desc* and *color* attributes. The desc attribute, short for 'description', describes the origin of the node and is one of: Study, Arm, Drug, Company, Insider. The color attribute contains a color code for the node in hexadecimal RGB format. It can be used with the *Custom Properties Mapper* in the yEd graph editor[27] to apply the color code as node color. This allows easy visual distinction between node types when inspecting the graph in yEd.

In the network, there are 16679 nodes and 22162 edges. The distribution of the different nodes is as follows: 561 drug nodes, 5152 trial nodes, 7288 arm nodes, 164 manufacturer nodes, 3514 insider.

The degree distribution for the various node types shows how the various parts of the network are connected. In the degree histogram for trial nodes (see figure 3), which only have outgoing edges to arm nodes, it is obvious that at ~ 3600, the vast majority of trials have only one arm. There are approximately 1500 trials with more than one arm, which will later serve as a criterion for establishing drug testing links.

In the out-degree histogram for insider nodes (figure 4), we can again see that most nodes only have one outgoing edge. Approximately 3200 insiders are only related to one company and thus do not link a company pair as 'friendly' while about 400 insiders belong to two companies or more.

[26] `http://graphml.graphdrawing.org/`
[27] Starting in Version 3.7.0. Available at `http://www.yworks.com/en/products_yed_about.html`

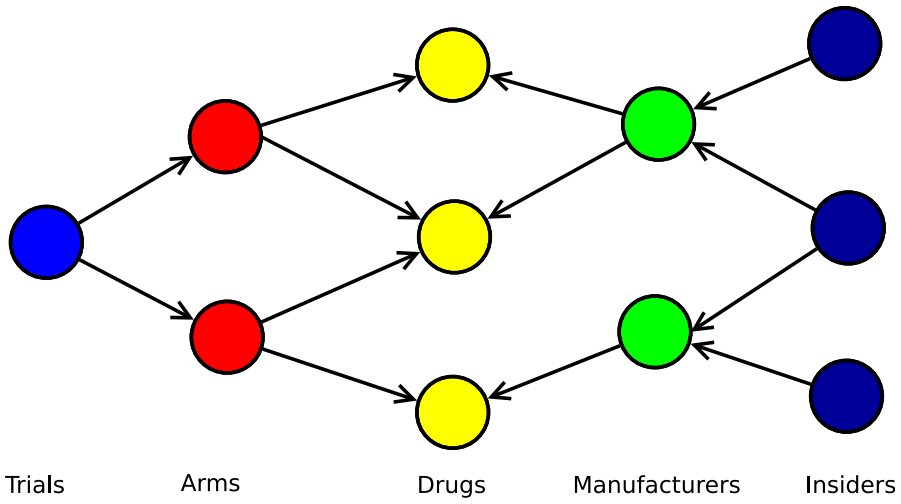

Figure 2: Network layout

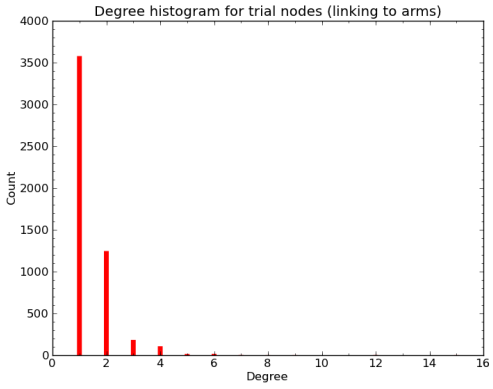

Figure 3: Degree distribution: links from trials to arms. Shows arm count per trial.

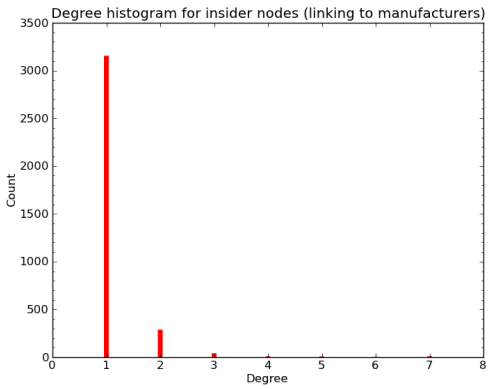

Figure 4: Degree distributions for outgoing links from insiders to manufacturers. Shows manufacturer count per insider.

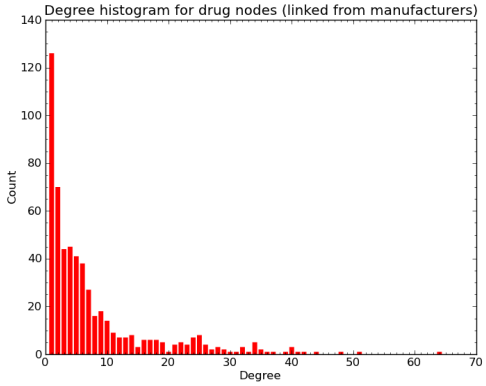

Figure 5: Degree distribution for drugs, linked from manufacturers. Shows manufacturer count count per drug.

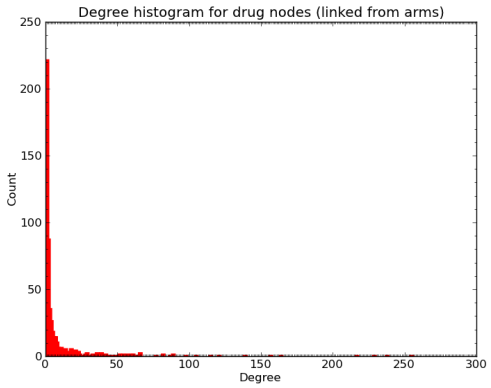

Figure 6: Degree distribution: incoming links to drugs from arms. Shows trial arms per drug.

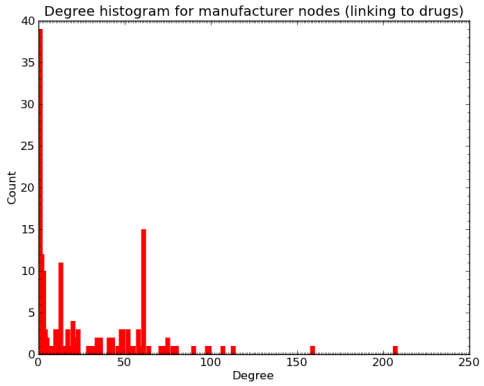

Figure 7: Degree distribution: manufacturer nodes linking to drugs. Shows drug count per manufacturer.

In the degree distribution histogram for drug nodes, counting incoming nodes from manufacturers, (figure 5), it is obvious that most drugs have only one manufacturers. Very few drugs have more than forty manufacturers, with one outlier having 64 manufacturers.

7. Analyzing the Network

We have now linked medical trials, drug manufacturers and insiders together to form a network. We will now proceed with network analytic methods to investigate our hypothesis.

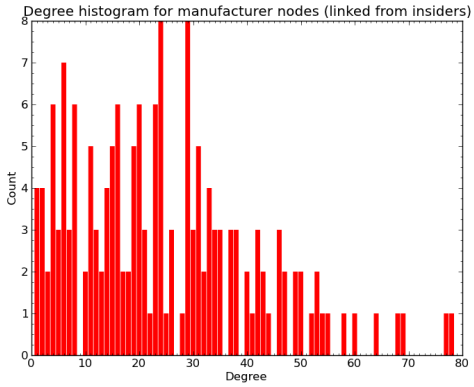

Figure 8: Degree distribution: incoming links to manufacturers from insiders. Shows insiders per manufacturer.

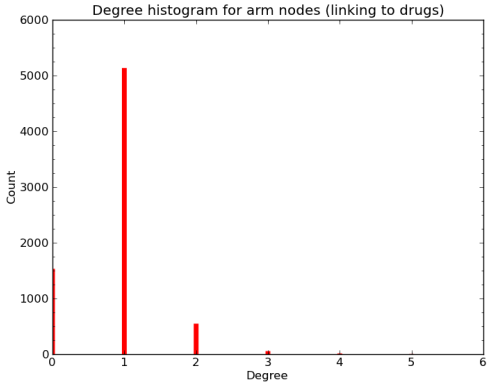

Figure 9: Degree distribution: arms linking to drugs. Shows drug count per arm.

7.1. One-Mode Projection

The analysis of our hypothesis basically consists of producing two different views on the set of companies, then comparing the views. One view is created by linking the companies based on their common insiders, the other view is created by linking the companies based on compared drugs.

In a network analytic approach, this is done using a one-mode projection (OMP) on a bipartite network. For the sub-network containing only insiders and companies, it is obvious that this network is bipartite. A few entities are both insiders and companies, as a company may buy shares of another company. This eliminates the bipartite-ness of the graph. For entities where both labels apply, we default to the company role as there are about ten times more insiders than companies. This step re-establishes the bipartite property of the graph.

The network between companies and drugs is bipartite as well; however, we also need trial and arm nodes to link compared drugs.

OMP on the insider graph For a traditional one-mode projection, it is sufficient to link any companies having at least one common insider. Assuming two companies A and B both have an insider X, we would consider A and B to the friends and connect them with an edge in the projection. To put it differently, if $cooc(A, B) > 0$, then A and B are connected in the projection. The resulting graph will henceforth be called *insider projection*.

OMP on the trial graph The portion of the network consisting of trials, trial arms and drugs is not handled as easily as it is not a bipartite network. If we reduce the network to drug and company node types, we get a bipartite network, but we lose the information which drugs are compared in medical trials. As such, it makes sense to define a more complex motif on the network and connect companies based on its occurrence. We start by defining a motif to capture pairs of compared drugs:

Given a trial T with at least two arms T_1 and T_2. Drugs x and y are considered *compared* if all of the following conditions hold:

- $x, y \in T$
- $T = \{T_1, T_2, T_3, ...\}$
- $x \neq y$, $i \neq j$
- $x \in T_i \wedge x \notin T_j$
- $y \in T_j \wedge y \notin T_i$

Some examples will make the conditions clearer. Consider a hypothetical study T for a new type of pain killer, named x. The trial is designed to compare the efficacy of x to pain killer y which is already approved and on the market. Accordingly, the trial has two arms T_1 and T_2, both consisting of a group of study subjects. Drug x is administered to

the subjects in group T_1, drug y is given to the subjects in group T_2. We can consider drugs x and y to be compared, as all of the conditions above hold.

As another example, consider a hypothetical study with the goal to find out if a combination of the pain killers x and y is more effective than x alone. Such a study can again be designed with two arms T_1 and T_2, each with a group of study subjects. Group T_1 receives the single pain killer x and group T_2 receives x and y as combined intervention. Which drugs are being compared here? For our purposes, we argue that no direct drug comparison is taking place. x is not being compared to y, but to a combination of x and y. If the combination of x and y performed better than x alone, sales of x would not necessarily decrease. As this work is based on the underlying hypothesis that pharma companies influence trials to steer revenue streams, we do not consider the drugs to be compared. This motivates the inclusion of the last two conditions $x \notin T_j$ and $y \notin T_i$ in the list above.

As a last example, consider a study comparing two combinations. x and y are assigned to T_1 while x and z are assigned to T_2. For this case, we can assume that y and z are being compared.

Given these pairings of compared drugs, we can easily look up their manufacturers. Given that drugs x and y both can have more than one manufacturer, the Cartesian product of the manufacturer sets for each drug is used to build the list of 'companies comparing their drugs'. Say x is manufactured by $\{a, b\}$ and y is manufactured by $\{b, c\}$, we would add $\{a, b\}$,$\{a, c\}$and$\{b, c\}$ to the list of 'companies comparing their drugs'. Note that we do not add $\{b, b\}$ as such a combination does not reflect competition between companies. Also note that the ordering of the company pairs in the Cartesian product does not matter, so we write these as sets. The one-mode projection is created by drawing edges between these companies.

Finally, the one-mode projection for the trial portion of the graph is created by adding edges between each pair of companies which compared their products. This perspective on the set of manufacturers will be referred to as *trial projection* or *drug projection*.

7.2. Analyzing drug comparison frequency between friends and strangers

We have now obtained two different perspectives on the set of companies. Each perspective is realized as a network, with related companies being connected with an edge. The insider projection features edges between friendly companies and the trial projection features edges between companies which compare their drugs in a clinical trial.

To test our hypothesis, we looked into how likely friendly companies are to compare their drugs, compared to the likelihood with which all companies are testing their drugs. To this end, we defined companies connected by an edge (path length 1) in the insider graph to be friends. Similarly, we defined companies connected by an edge in the trial graph to be companies which compare their drugs.

We first assessed the likelihood for random pairs of companies to compare their drugs in at least one trial and its variance by sampling random pairs of companies from the trial graph with a bootstrap method: we took 1000 samples of n pairs; n being the number of friendly company pairs. This data was plotted as box plots, such as figure 10. The

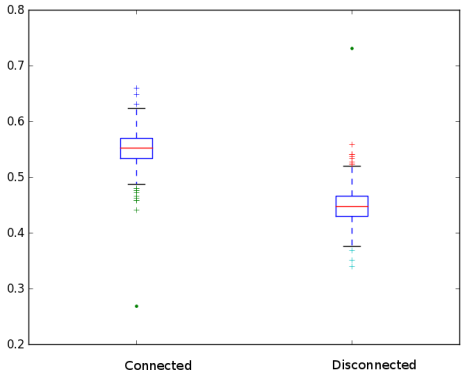

Figure 10: Trial p=all, Insider p=all

red line is the median of the data, and the box extends to the first and third quartile of the data. The whiskers extend to the most extreme datums still within 1.5x inner quartile range, starting from the first and third quartile respectively (Q1 - 1.5xIQR and Q3 + 1.5xIQR). Assuming a normal distribution, the whiskers encompass 99.3% of the data range (up to a standard deviation of $\sigma = 2.698$). Therefore, any outliers outside the whiskers are statistically significant at $\alpha = 0.7\%$.

We then estimated the frequency with which friendly companies compare their drugs in at least one trial by counting how many of the friendly company pairs are represented in the edge set of the trial graph. This frequency is shown as green dot in the box plots.

As can be seen in figure 10, approximately 55% of the companies sampled uniformly at random are connected by an edge. Conversely, 45% are not connected. If we limit the sample to friendly company pairs, about 73% of these friendly companies are not connected in the trial graph. Given that the whiskers for the box plot for disconnected pairs extend to be about 53%, we can safely say that 27% is a statistically significant observation.

This means that, at least on our limited data set, friendly companies are less likely to compare their drugs in at least one clinical trial than random pairs of companies.

8. Test against null model

Based on the traditional one-mode projection, we have found evidence supporting our hypothesis. However, the method we employed to create the projections is far from ideal. The traditional method simply links two companies once there is at least one common insider. Similarly, the criteria to establish the *compared* relationship between drugs only requires a single trial. However, the strength of the relationship is ignored completely. If companies A and B have two common insiders, while B and C share five common insiders, which companies have a tighter friendship? This information is relevant when testing our hypothesis as we would expect closer friends to be able to exert more control

over each other.

It follows that we should use the number of common insiders or the number of trials for a drug pair as some kind of weight to assess the relevance of the link in question. This approach captures more information on the network, but still has some problems. Consider two companies A and B which share five insiders. Assume that each company has 30 insiders in total, of which only five insiders belong to both companies. Consider a second pair of companies C and D sharing five insiders. However, C and D each only have six insiders in total. Clearly, the link established by five insiders between A and B is weaker than the link established by five insiders between C and D. If a company has many insiders, the individual insider's influence bears less weight than in a company featuring few insiders.

Simple Independence Model As laid out above, it is prudent to consider the degrees of the individual company nodes when assessing relationship strength. We thus consider the Simple Independence Model, which compares the observed value for a node pair, i.e. the number of common insiders or the number of trials for a drug pair, against the expected value based on the model.

This model originally stems from the field of *market basket analysis*(Zweig, 2010). The original question there is: if a customer buys a product a, what other products are they likely to buy? This information is e.g. useful for recommendation systems. In this scenario, we have a network consisting of products and customers. If a consumer buys a product, an edge is added to the network between the consumer and the product. Based on this network, the goal is to find out which products are bought together more frequently than pure chance would suggest it.

In this model, the likelihood of an user buying a single product a, the so-called *support*, is estimated by the relative frequency with which a is bought. Since a sale of a is modeled in the network as an edge between a and the customer, the support for a is defined as $supp(a) = \frac{deg(a)}{r}$, with r being the total number of edges in the network.

Based on the estimated likelihood for a single product, the expected value for selling two products a, b together if a and b are independent is given by: $supp(a) \cdot supp(b) = \frac{deg(a)}{r} \cdot \frac{deg(b)}{r}$

The actually observed frequency with which a and b are sold together is estimated using the co-occurrence measure *cooc*. This measure counts how many customers bought products a and b together. To obtain the relative frequency, *cooc* is normalized against the number of all edges r to form the support measure $supp(a, b) = \frac{cooc(a,b)}{r}$.

If sales of a are independent from sales of b, then we would expect the value for $supp(a, b)$ to be same as the expected value $supp(a) \cdot supp(b)$. If $supp(a, b)$ is larger than the expected value, we know that a and b are bought together more often than pure chance would suggest. Thus, b is a possible recommendation candidate for customers who buy a and vice versa. If desired, ranking between recommendations is done based on the difference between expected and observed value in the original paper.

The underlying model (Simple Independence Model, SIM) thus assumes that each company is equally influenced by an insider with probability (or weight) $deg(v)/r$. To put

it differently, for a given company, each insider takes influence with the same probability, which is described by an uniform distribution.

Fixed Degree Sequence Model Using an appropriate null model allows us to compare the importance of a company pair against the expected value and is thus very valuable. However, the Simple Independence Model still fails to take into account the degree of the insider nodes. Consider an insider X which belongs to 20 companies. If companies A and B share this insider X, then we would create an edge between A and B with weight 1. However, consider an insider Y which only belongs to 2 companies, and these two companies are A and B. Surely this insider Y is more important than insider X, given that Y may have a special interest in these two companies. In a sense, insider X is less sensitive to an individual company failing since his investment is probably more distributed.

For the reasons outlined above, it is prudent to use a better model than SIM to assess the significance of an observed *support* value. Zweig (2010)[28]describes the*fixed degree sequence model*(FDSM) which takes into account the degree distribution of both sides of the network, i.e. the degree distribution for companies and insiders. As defined by Zweig (2010):

> Given a graph $G = (V = \{V1, V2\}, E)$ and the corresponding degree sequences L and R, we define $G(L, R)$ as the set of all bipartite graphs with the same degree sequences (and no multi-edges).

Zweig notes that the SIM is ill-suited for graphs with a strongly skewed degree distribution, i.e. degree distributions with a long tail. As we can see in figures 7 and 5, the degree distributions are indeed skewed, further substantiating the need for FDSM which models this scenario better.

As the set $G(L, R)$ is very large even for small graphs, it is unfeasible to compute the average co-occurrence (support) by enumerating all graphs in the set. However, sampling from $G(L, R)$ is possible by means of a simple Markov process:

> Starting from a Graph in $G(L, R)$, in each step two edges (v, w) and (x, y) are drawn uniformly at random from the set of all edges. If a swap of these edges does not lead to a multi-edge, i.e., if neither (v, y) nor (x, w) is already in E, (v, w) and (x, y) are replaced by(v, y) and (x, w).

It should be noted that G itself is in $G(L, R)$ and can thus be used as a starting point for the Markov process. Given a sufficiently large number of random walk steps, the swap operation is guaranteed to end at each graph in $G(L, R)$ with the same probability, as shown in Cobb and Chen (2003). In fact, the algorithm we use is a specialization of algorithm 22H in (Cobb and Chen, 2003).

Quite like the SIM, the method proposed by Zweig (2010) compares the expected value for a given company pair with the observed value. To this end, a sufficiently large

[28]We would like to note that an extended version of this article was published in 'Social Network Analysis and Mining'(Zweig and Kaufmann, 2011)

number of graphs from $G(L, R)$ is sampled and the *cooc* for all node pairs is computed in each graph. The expected $cooc_{FDSM}$ value for a company pair a, b is then given by the arithmetic mean of all values $cooc(a, b)$ seen during the sampling. Again, the observed *cooc* from the original graph is compared with the expected $cooci_{FDSM}$ to assess the importance of the observation. The pairs of companies may then be ranked according to their observed *cooc* value, corrected by the expected $cooc_{FDSM}$ value.[29] However, even this approach is not optimal, as noted by Zweig in private conversation. The difference between the observed value and the expected value is not enough, as we do not know if this difference is statistically significant[30]. Consider two hypothetical distributions for two given company pairs (a, b) and (c, d) for *cooc*, obtained by sampling from $G(L, R)$. In a distribution with a large standard deviation, even a seemingly big difference between observed *cooc* and expected $cooc_{FDSM}$ value may be statistically insignificant. On the other hand, assuming a distribution with a very small standard deviation, even a small difference from the mean may be statistically significant. It is thus prudent to note how rarely we encounter a value during sampling that is at least as extreme as the observed value in the original graph for a given node pair. For a sample size n, we let $cooc_i(a, b)$ denote the co-occurrence for nodes a, b in the i-th graph sampled from $G(L, R)$, where i is a value from 1 to n. We further define $o_i(a, b) = \begin{cases} 0 & \text{if } cooc(a, b) \leq cooc_i(a, b) \\ 1 & \text{if } cooc(a, b) > cooc_i(a, b) \end{cases}$. The p-value for a, b is then defined as $p(a, b) = \frac{\sum_{i=1}^{n} o_i(a,b)}{n}$. By counting how often the observed value exceeds the sampled value, we can estimate the p-value.

8.1. A null model for the trial graph

We have so far described how to test observations on our network against a suitable random graph model, namely the FDSM. However, the FDSM as described by Zweig (2010) is only defined on bipartite networks. This does not pose a problem to us for the insider projection, as the sub-network consisting of insiders and companies is indeed bipartite. However, the trial projection is based on a complex network motif considering trials, arms, drugs and manufacturers and thus cannot be classified as bipartite anymore. Fortunately, adapting the FDSM for bipartite graphs is feasible if we ask ourselves which observations on the network we would like to test and which observations we would like to consider fixed. We can decompose the trial portion of the graph into three distinct relationship categories:

- Which arms does a clinical trial have?
- Which drugs are tested in a trial arm?
- Which companies produce a drug?

[29] This measure is called *leverage* in Zweig (2010)

[30] It is possible that the distribution of sampled *cooc* values looks similar over all company pairs, thus making the difference between *cooc* and $cooc_{FDSM}$ a good measure, but it has yet to be proven.

Randomizing the relationship between trials and arms or arms and drugs is not useful, as this structure must be considered fixed and given by the clinical trial data. As we are interested in significant relationships between drugs and their manufacturers, it is adequate to swap the corresponding edge subset to find out which links are merely dictated by the structure of the network and which are actually significant, i.e. happening more often than pure chance. Given that the sub-network consisting of manufacturer and drug nodes is bipartite, we simply use the swap operation defined above.

Now that we have adapted the FDSM to the trial portion of the graph, we need a measure for relationship strength between companies a,b in the trial network. Basically, we need a counterpart for the co-occurrence measure employed for the insider projection which will work in the more complex trial graph. To this end, we define $motif(a, b)$ to be the frequency with which the motif defined in 7.1 occurs between a and b. In other words, we count how often companies a and b compare their drugs. We then sample random graphs from $G(L, R)$ as described above and compute for all company pairs a,b the measure $motif(a, b)$ for each sampled graph. The p-value is then recorded for each pair by counting how often the sampled $motif(a, b)$ was was high as the value observed on the original network, divided by the number of samples for normalization.

8.2. Sampling parameters

For each of the two projections, 500 random graphs were generated to calculate the p-value for the node pairs. Each random graph was created by doing $m \times \log(m)$ swaps as suggested by Zweig (2010), with m being the number of edges per graph. The subnetwork consisting of trials, drugs and manufacturers has 18146 edges and the subnetwork consisting of manufacturers and insiders graph has 4016 edges.

8.3. Pruning the network

We now have a way of assigning p-values to pairs of companies for both projections. We now need a way to select only the best, i.e. the most significant edges. Traditionally, one would choose a significance threshold such as $\alpha = 0.05$ or $\alpha = 0.10$ and discard any node pairs whose p-value is larger than α. However, an adaption of the work done by Zahoranszky et al. (2009) provides a more interesting perspective which considers network information when choosing the threshold.

As we are interested in finding friends, and perhaps even groups of friends, the concept of cliques becomes of interest. How close a graph is to being a clique is measured by the clustering coefficient. We are basically interested in obtaining a high clustering coefficient, to maintain tight groups of friends or groups of companies who often compare their drugs, while simultaneously only adding significant edges.

To find the interesting points at which cluster coefficient is relatively high while restricting the set of edges to the most significant node pairs, we employ the following algorithm to create a plot of p-value threshold versus clustering coefficient:

Starting with the empty graph, we gradually add the most significant edges of an one-mode projection by p-value in ascending order, i.e. least significant edges will be

P-Value	Clustering coefficient	Edge count
0.0000	0.4087	103.0000
0.0100	0.3173	129.0000
0.0220	0.3088	133.0000
0.0600	0.2438	160.0000
0.0920	0.2715	201.0000
0.1000	0.2711	212.0000
0.1280	0.2818	236.0000

Table 4: P-value cut-offs for insider projection

added last. Edges with the same p-value, so-called ties, are added at once. After each iteration, we compute the clustering coefficient[31] for the graph. The cluster coefficient at each p-value is recorded in a plot and later reviewed manually. In these plots, we focus on peaks where the cluster coefficient has a local maximum.

Adding all edges in one-mode projections leads to very dense networks and the clustering coefficient for these will always be high. Thus, clustering coefficient is used to perform a trade-off operation: if we desire to only add the most significant edges, when do we have to stop to get the most dense network? This trade-off manifests itself in the peaks in the graphs, after which the average clustering coefficient often drops off sharply. Based on these peaks, we prune the projections by discarding any node pairs whose p-value is higher than the newly found thresholds.

Projection from Insiders to Companies Figure 11 shows the cluster coefficient for the projection from the insider graph onto the company graph at various p-value cut-offs. Figure 11 shows the plot of p-value against the cluster coefficient. It shows that at p-value 0, the cluster coefficient is highest at around 0.4087. We can interpret this to mean that the most exceptional company pairs create a very clustered network. Further local maxima can be found at various points up to 0.1280, after which the cluster coefficient for the pruned graph sharply declines. At p-value approximately 0.1600, the cluster coefficient rises again and then stabilizes around 0.2500. It is obvious from the plot that adding edges with a p-value higher than 0.1280 is not useful as no interesting maxima can be found. Instead, the cluster coefficient is stable at an uninteresting level. This can be explained with the distribution of p-values over the company pairs. Looking at figure 12, we can see that only few edges (approximately 10 out of 280) have p-values higher than 0.2500, which is the point where the cluster coefficient is stagnating around 0.1280 in figure 11. Only few edges are added at each threshold after that point, which is why the clustering coefficient does not change much.

The p-value cut-offs we have chosen and the accompanying network sizes can be found in table 4.

[31] From the NetworkX package: nx.algorithms.cluster.average_clustering()

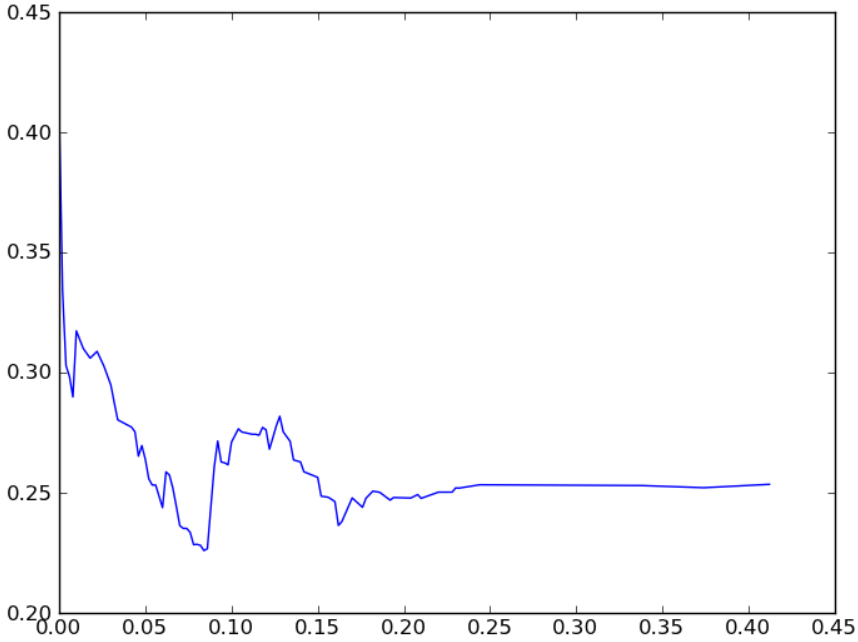

Figure 11: P-value (x-axis) plotted against cluster coefficient for insider projection

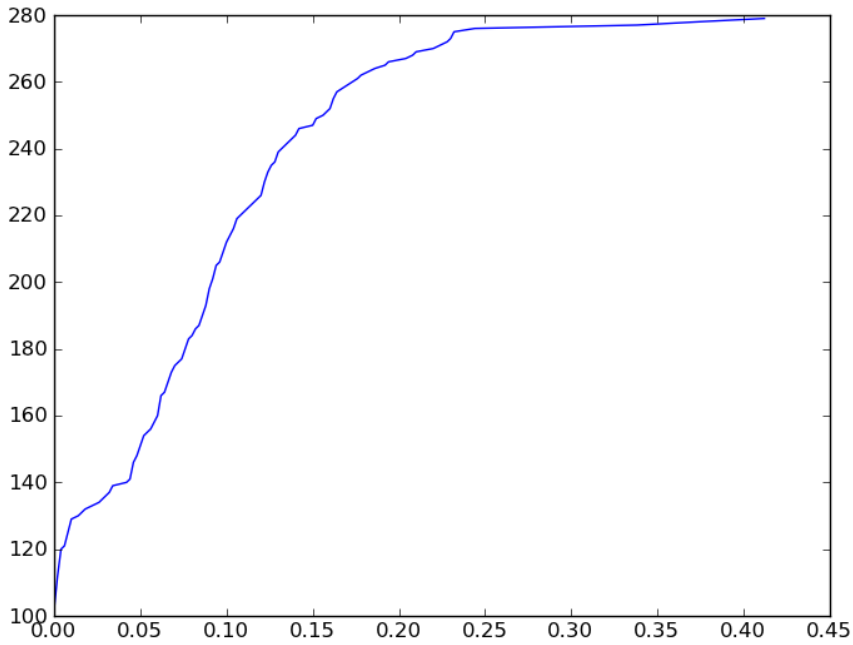

Figure 12: P-value (x-axis) plotted against edge count for insider projection

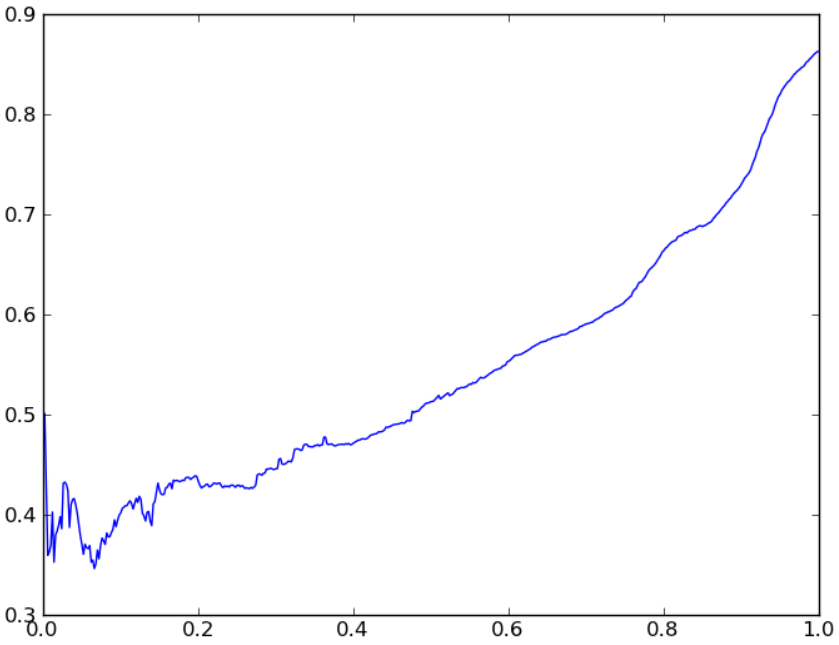

Figure 13: P-value (x-axis) plotted against cluster coefficient for trial projection graph

Projection from Trials, Arms and Drugs to Companies Looking at the plot of p-value against cluster coefficient for the drug projection, it is obvious that the network with the highest clustering coefficient is the network which retains all company pairs. For lower p-values, i.e. more significant edges, we again see a big peak close a p-value of zero. In this case, pruning the network at a p-value of 0.0200 yields a clustering coefficient of 0.5011, encompassing the first 72 edges (out of almost 5000 company pairs). We listed further interesting local maxima in table 5, such as another interesting peak at 0.4326. While the clustering coefficient is rising steadily, we deemed the accompanying high p-values as not sufficiently significant.

Plotting p-value against resulting edge counts in figure 14 shows that p-values are distributed rather uniformly, unlike the edge count graph for the insider network (figure 12) .

The fact that the average clustering coefficient continues to rise when adding more edges simply reflects the nature of the average clustering coefficient when used on naturally dense networks.

The p-value cut-offs we have chosen and the accompanying network sizes can be found in table 5.

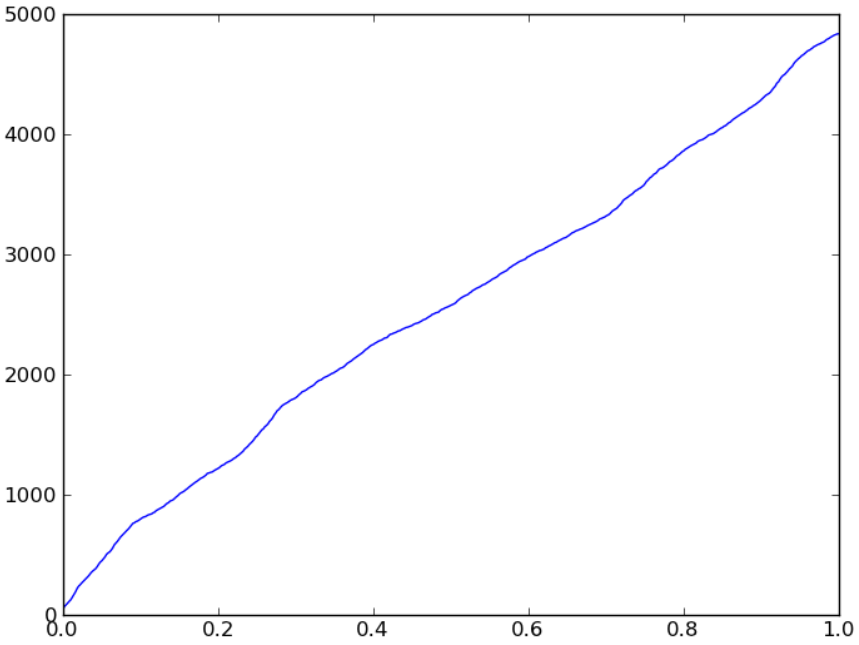

Figure 14: P-Value (x-axis) plotted against edge count for trial projection

P-Value	Clustering coefficient	Edge count
0.0020	0.5011	72
0.0120	0.4026	194
0.0280	0.4326	293
0.0400	0.4162	374
0.1480	0.4316	993
0.2880	0.4457	1764
0.3540	0.4701	2036

Table 5: P-value cut-offs for drug projection

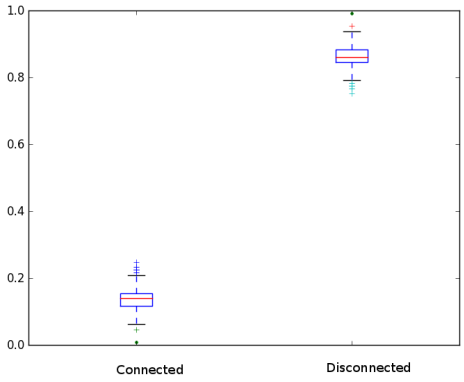

Figure 15: Trial p=0.012, Insider p=0.01

8.4. Analysis

We have chosen graphs at approximately similar p-value thresholds for comparison as well as some more extreme combinations. For combinations shown in figures 15, 16 and 17 at cutoffs $p_t = 0.012$, $p_i = 0.01$; $p_t = 0.04$, $p_i = 0.06$ and $p_t 0.028$,$p_i = 0.022$, the box plots clearly show statistical significance. When sampling pairs of companies uniformly at random, approximately 10% of these companies will be linked by an edge. Conversely, about 90% are not be connected. If we limit the sample to friendly company pairs, i.e. companies connected by an edge in the insider projection, about 99% of these friendly companies are disconnected in the trial graph. We therefore have evidence that friendly pharma companies compare their products significantly less than strangers.

At p-value thresholds of $p_t = 0.288$,$p_i = 0.01$ (figure 18) and $p_t = 0.354$,$p_i = 0.092$ (figure 19), the frequency differences become statistically insignificant. However, the same trends can be observed.

For $p_t = 0.354$,$p_i = 0.01$ (figure 20), the difference once again becomes significant. The same observation is made on the unpruned graph (figure 10), where the difference is significant even on the graphs where all edges are included.

8.5. Node distance agreement between projections

Early in the research process, we evaluated another approach to test our hypothesis. We finally abandoned this approach due to problems described below, but we would like to present our idea and its shortcomings for the sake of completeness.

As an additional way to test our hypothesis, we decided to look at the agreement for shortest path distance between pairs for both projections. Here, we use the path length of the shortest path between two companies as a measure for their closeness. Since for each graph, the path length, or the distance, essentially defines an ordering on the set of node pairs, we can measure how much these two orderings on the set of node pairs agree with each other.

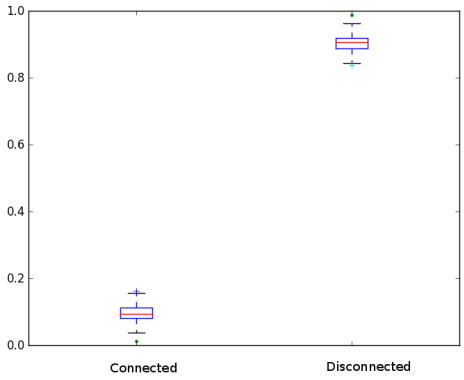

Figure 16: Trial p=0.04, Insider p=0.06

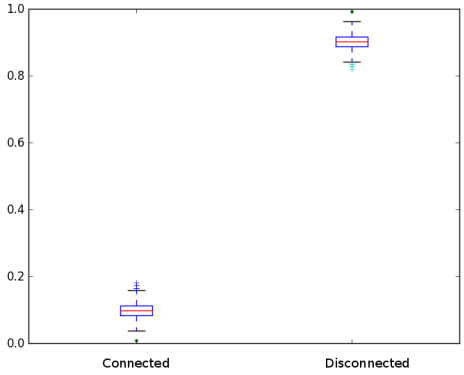

Figure 17: Trial p=0.028, Insider p=0.022

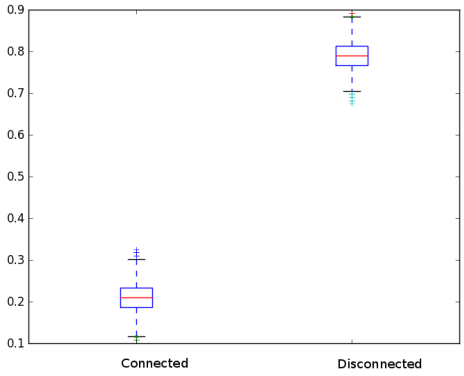

Figure 18: Trial p=0.288, Insider p=0.01

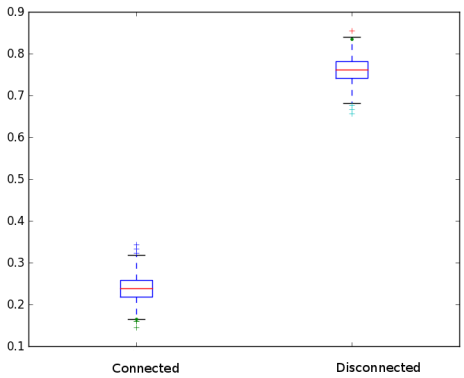

Figure 19: Trial p=0.354, Insider p=0.092

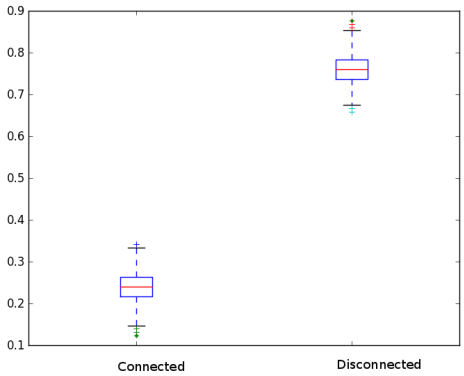

Figure 20: Trial p=0.354, Insider p=0.01

Kendall's tau is such a rank correlation coefficient(Kendall, 1962). To compute Kendall's tau for the distance rankings, we first compute for each company pair the respective distance in trial and insider projections. From this distance, pairs $(d_t(A,B), d_i(A,B))$ are created where $d_t(A,B)$ is the distance computed in the trial projection for a company pair (A,B) and $d_i(A,B)$ is the distance computed in the insider projection for the same company pair (A,B). Given two pairs $(d_t(A,B), d_i(A,B))$ and $(d_t(C,D), d_i(C,D)$, we determine whether this pair is concordant or discordant, e.g. whether the relative ordering is the same or not. The relative ordering is the same if $d_t(A,B) < d_i(A,B) \wedge d_t(C,D) < d_i(C,D)$ or $d_t(A,B) > d_i(A,B) \wedge d_t(C,D) > d_i(C,D)$.

We do this for all possible pairs and then subtract the number of discordant rank item pairs from the number of concordant rank item pairs among the orderings and normalize this against the total possible number of pairs. Thus, a Kendall's tau value of 1 indicates complete agreement while a value of -1 indicates complete disagreement. Given our hypothesis that friendly companies do not compare their drugs, we expect a low value for Kendall's tau, as the distance for a given company pair should be high in one projection and low in the other projection.

Problems with this approach When comparing two rankings with Kendall's tau, both ranking must have the same dimensions. However, this is not the case for our problem as we compute the shortest path length for all possible node pairs. For disconnected node pairs, the shortest path is not defined. As we can see in 4 and 5, the difference in edge count between the insider projection and the trial projection can be substantial. This results in fewer connected node pairs for the insider projection, leading to a shorter ranking list for this projection.

To rectify this problem, only the common subset of connected nodes between both projections was included when calculating the rank, i.e. companies A and B are only included in the ranking if a path exists between A and B in both the trial projection and the insider projection. This ensures that both ranks are of equal length.

However, this also needlessly discards paths which only exist in one projection and introduces a bias towards connected nodes. If a company pair is connected with, suppose, path length 1 in the trial projection and not connected at all in the insider projection, this would indicate a strong disconnect and strengthen our hypothesis. Indeed, preliminary analysis has shown that Kendall's tau usually lies between 0 and 1, indicating positive correlation. This is not surprising given that by omitting disconnected node pairs, we actually discard the data points which could support our hypothesis.

As a second problem, the notion of distance in the trial projection is not meaningfully defined. If two manufacturers X and Z are connected with distance 2, then this is the result of two trials: a drug produced by manufacturer X is tested against a drug manufactured by a company Y, and in a second trial, a drug made by Y is tested against a drug sold by Z. This scenario is not relevant to us as we are primarily interested in company comparing their drugs directly.

Given these problems, we decided to abandon this approach and focus on the analysis described in Section 7.2.

Part III.
Conclusion

9. Suggestions on current Open Data Efforts

One of the reasons for governments to release open data is to enable better analysis by the scientific community. To this extent, we identified two big problems when working with the data. First, the quality of the data itself often is not very good. In case of the ClinicalTrials.gov data set, the data is only semi-structured, i.e. a lot of information is hidden in natural language text and not easily machine-readable. Consider the intervention_name tag, which only consists of a string such as "manidipine 10 mg + delapril 30 mg". It would be much more useful to have a more fine-granular description of the intervention, with separate tags for dosage and administration frequency. For interventions composed of multiple drugs, a clear distinction between individual drugs must take place. As an example how this data could have been better represented for our purposes, consider the following XML representation:

```
<intervention>
<drugs>
<drug>
<drug_name>manidipine</drug_name>
<drug_dosage unit="mg">10</drug_dosage>
<drug_administration>oral</drug_administration>
<drug_frequency>daily</drug_frequency>
</drug>
<drug>
<drug_name>delapril</drug_name>
<drug_dosage unit="mg">30</drug_dosage>
<drug_administration>oral</drug_administration>
<drug_frequency>daily</drug_frequency>
</drug>
<arm_group_label>1</arm_group_label>
</drugs>
</intervention>
```

The proposed XML schema would have made it easy to extract individual drugs (and disregard dosage information) and immediately see if they are assigned to different arm groups. Note that designing a proper schema or an ontology for real-world data never is an easy task as corner cases will always make the design more complicated. However, not all data needs to be properly structured; even some structure such as individual drug lists would have benefited us greatly.

A similar proposal can be made for the description of the study design. Consider "Allocation: Randomized, Endpoint Classification: Safety/Efficacy Study, Intervention

Model: Parallel Assignment, Masking: Open Label, Primary Purpose: Treatment", which is the study_design tag for NCT00450580. The structure underlying the classification of medical trials should be easily turned into an XML schema, providing structured access to this information. Yet, if a researcher were to filter or compare trials based on the "intervention model", they would have to deal with extracting the proper information from natural language.

The case we would like to make here is that with reasonable manual or automatic effort, the usefulness of the data could be vastly improved. It is useful to properly structure the data at the source instead of relying on data consumers to extract desired information, as this leads to duplication of effort. As stated before, complete machine readability is unlikely to be achieved, but some structure goes a long way.

Lacking normalization of named entities Another problem is the lacking normalization of the data. As described in Section 6.2, linking records from two data sources is a daunting task. We propose that an URI (Uniform Resource Identifier) is assigned to each entity involved by the data sources. In the case of EDGAR, the URI could be based on the Central Index Key. For drugs, the DrugBank ID is available.

DrugBank listings use varying strings to refer to the same entity. The same issue applies to ClinicalTrials.gov interventions and the drugs contained therein. Even when restricting data usage to a single source, e.g. using only ClinicalTrials.gov or DrugBank data, it is impossible to query the data source for all entries related to a certain entity. Listing all drugs manufactured by a given company or listing all manufacturers for a given drug becomes difficult as the entity mentions are not normalized.

However, even if each data source employs their own scheme to assign unique identifiers to entities, it will still be hard to create cross-source record links. While maintaining global databases of URIs is certainly not feasible in the near future, unique identifiers already exist for the data at hand and can simply be re-used, such as the DrugBank ID and the EDGAR CIK. As an alternative, each drug likely must be registered with the NIH (National Institutes for Health) which likely assigns an internal ID. Similarly, companies most likely have a tax id which uniquely identifies them and is known to the public.

In the case of the ClinicalTrials.gov data set, trial data is submitted electronically via the protocol registration system[32]. If the onus is on the individual trial registrants to provide the unique identifier for the drugs, be it an URI or simply the DrugBank or PubChem identifier, no additional load is placed on the data source itself. The additional load on the trial registrant is negligible.

A similar process may be proposed for the DrugBank manufacturer information. While the actual source of manufacturer information is unknown to us and an inquiry has not been answered, we speculate that the data is based on actual forms submitted by the manufacturers. Again, if the manufacturer simply submits their identifier such as their tax number, or if applicable, the CIK, this problem is helped immensely. Even if no unique identifier per se is available, additional information such as the exact company

[32] `http://prsinfo.clinicaltrials.gov/`

name, founding date and founding location, would help greatly in creating a composite key as identifier.

10. Future Work

10.1. Network creation

Data sources The initial data sources provide more information than what is currently being used. For example, drug meta data could be used to influence network generation or the network null model. DrugBank data lists drug categories, classes and drug indications (i.e. for which diseases a given drug is used as treatment) as additional information. This could be used to find out which drugs are more likely to be compared. Additional information on pharmacological action, pharmacological classification and chemical classification is available from PubChem[33], which can be accessed via DrugBank. ClinicalTrials.gov also lists sponsorship information for clinical trials. It would be of interest to limit the data set to industry-sponsored studies to see if a pattern different from what we have seen so far emerges.

If we assume that only drugs for similar indications will be compared, it will be easier to see if some companies are not testing their drugs as intensively as other companies. DrugBank also lists pricing information, which might serve as an indication how much a drug should not be damaged. There is less damage being done between friends if a less expensive drug is pushed out of the market. However, sales volume in general would have to be taken into account as well.

It should also be noted that the network of pharma companies is not necessarily complete as not all pharma companies are listed by SIC 2834, if they're listed in EDGAR at all. Additional categories such as '8731' ("SERVICES-COMMERCIAL PHYSICAL & BIOLOGICAL RESEARCH") might prove to contain companies of interest.

Insider definition As laid out in Section 5.3, the quality of the insider lists extracted from the ownership reports in EDGAR is questionable. For future research, the origin of the data should be clarified and, if necessary, only current insiders should be considered.

Entity Resolution While inter-source named entity resolution works reasonably well with an F-Score of about 0.8, multiple mentions within a single source are not recognized.

Consider the case of "ALLERGAN INC" (CIK 0000850693), "ALLERGAN LIGAND RETINOID THERAPEUTICS INC" (CIK 0000934592) and "ALLERGAN SPECIALTY THERAPEUTICS INC" (CIK 0001049711). Again, these strings refer to separate legal entities, but as they are very closely related we would like to consider them one single entity to create a denser network. These multiple mentions are not a single incident; consider further examples such as "ANDRX CORP" (CIK 0000911755) and "ANDRX CORP /DE/" (CIK 0001123337). Another example is listed in table 2.

[33] `http://pubchem.ncbi.nlm.nih.gov/`

A useful tool for record linkage might be DDupe[34]. DDupe is an interactive tool using data mining algorithms with user interaction to resolve potential duplicate pairs. DDupe works on relational data, thus incorporating rich network information (Kang et al., 2008)

Another way to improve record linkage would be improved string similarity metrics. For the linking of company names between EDGAR and DrugBank, we employed the Levensthein distance as string similarity measure. Additional string similarity measures are described in literature, such as the ones implemented by the SimMetrics software package[35], which might be used to cluster company mentions together into entity clusters. The authors of the SecondString software package provide an evaluation of the string similarity measures implemented in their package for record linkage tasks (Cohen et al., 2003), but do not solve the problem of finding appropriate cut-off thresholds for clustering to avoid introducing false positives.

It should be noted that the LODD project (Linking Open Drug Data)[36], part of the larger Linked Open Data effort, is facing similar challenges in resolving records in creating their linked data set. We mentioned the LinkedCT and the DrugBank RDF data set in Section 5.5. The wiki page[37] for the LODD project describes their data interlinking effort, but unfortunately does not provide an evaluation of the methods employed. As future work, it would also be interesting to evaluate the tools used by LODD with the evaluation sets created for this thesis. This requires data conversion work as the LODD tools (Linquer[38], Silk[39]) only work with RDF data.

Obeying temporal information In our network building process, we disregard temporal information. We do not consider the date of a clinical study, nor do we consider the date an insider enters or leaves a company. Given that we consider drug trials from 2006 to 2010, many changes in the insider might happen in the same time span. Drug trials might happen between companies before they are linked by a common insider. Conversely, insiders might have left companies before a significant reduction in drug tests for these companies happened later in time. Given these scenarios, it is important to consider temporal information in future work.

10.2. Considering insider intentions

As we have shown, friendly companies compare their drugs less often than random pairs of companies. We have proposed two possible explanations for this observation. The first explanation is that insiders influence the drug selection in studies to prevent damage of the revenue stream between friendly companies. The second, more harmless, explanation, is that insiders actively choose companies which are unlikely to compare their drugs, e.g. because they invest in different branches of the pharma industry.

[34] http://www.cs.umd.edu/projects/linqs/ddupe/
[35] http://sourceforge.net/projects/simmetrics/
[36] http://www.w3.org/wiki/HCLSIG/LODD
[37] http://www.w3.org/wiki/HCLSIG/LODD/Interlinking
[38] http://dblab.cs.toronto.edu/project/linquer/
[39] http://www4.wiwiss.fu-berlin.de/bizer/silk/

For this reason, future research should aim to resolve this question. In an approach similar to the analysis in this work, a network could be built linking companies, their drugs and the drug categories as provided by DrugBank. The one-mode projection on the set of companies would then link companies competing in the same branches. Given the existing insider projection, the question could be analyzed similar to Section 7.2.

10.3. Analyzing the path length distribution

During our analysis of the data, we have observed another interesting pattern.

Starting with the set of immediately connected company pairs in the trial projection, we sampled the distance of these nodes in the right graph. We compare this distance histogram with the distance histogram over all company pairs $V \times V$ in the insider projection.

As we can see in figure 21, about 20% of the node pairs from $V \times V$ of the insider projection are not connected. If we restrict ourselves to the subset of node pairs with path length 1 in the trial projection, i.e. companies which directly compare their drugs, we see that about 62% of these companies are not connected at all in the insider graph. This shows us that companies which compare their drugs very often do not have common insiders; statistically significant more often than for random pairs of companies.

When only considering the subset of connected company pairs, as seen in figure 22 for the unpruned projections, we see another interesting pattern: the distributions have a different shape. Even considering the fact that close companies from the trial projection are very often not connected in the insider projection, we might still expect the distance for the remaining connected edges to be similarly distributed compared to the random node pairs. However, we see distances 1, 2 and 3 happening significantly more often than baseline.

This indicates that companies which compare their drugs tend to be friendlier than random pairs of companies, if they are friendly at all. Overall, 70% of company pairs comparing their drugs are not friends, compared to a baseline of approximately 19%. However, the remaining 30% tend to be friendlier than average.

We do not have an explanation for the observed pattern, but we believe that further investigation will provide useful insight in the relationships of drug manufacturers and clinical research.

11. Results

After extracting information from various data sources, linking the entities contained therein and subsequently building a network between clinical trials, drug manufacturers and their insiders and then filtering out insignificant links, we have indeed found evidence validating our hypothesis. Friendly companies, linked by common insiders, have a significantly lower incidence of comparing their drugs than random pairs of companies. This might stem from the fact that insiders do not want their companies to compete, as this could result in a loss of revenue. As another explanation for our observation, it is possible that investors try to diversify their investments across a broad spectrum of

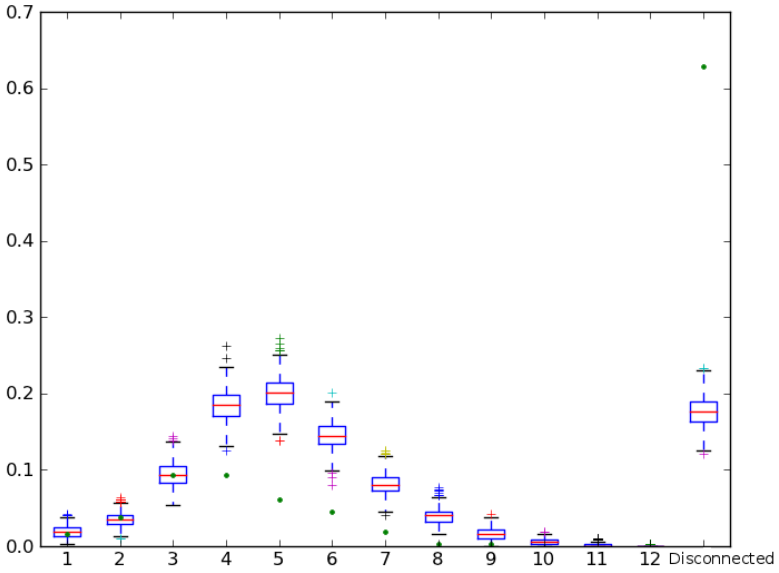

Figure 21: Distance histogram. Projection pruning thresholds: Insider 0.092, Company 0.04

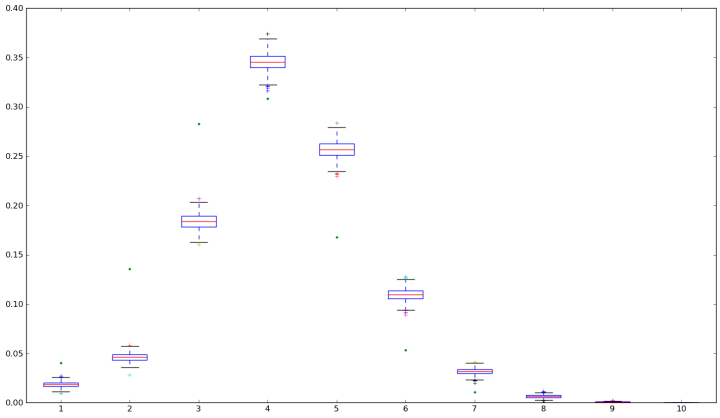

Figure 22: Distance histogram. Considering all insider pairs, all company pairs, only connected nodes.

pharma companies, thus selecting companies for investment which are very unlikely to compare their products. Future work will have to resolve this question.

We have also presented suggestions to improve the usefulness of current Open Data efforts.

12. Bibliography

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *In 6th Int'l Semantic Web Conference, Busan, Korea*, pages 11–15. Springer, 2007.

Justin E. Bekelman, Yan Li, and Cary P. Gross. Scope and impact of financial conflicts of interest in biomedical research. *JAMA: The Journal of the American Medical Association*, 289(4):454–465, 2003. doi: 10.1001/jama.289.4.454. URL `http://jama.ama-assn.org/content/289/4/454.abstract`.

Troyen A. Brennan, David J. Rothman, Linda Blank, David Blumenthal, Susan C. Chimonas, Jordan J. Cohen, Janlori Goldman, Jerome P. Kassirer, Harry Kimball, James Naughton, and Neil Smelser. Health industry practices that create conflicts of interest. *JAMA: The Journal of the American Medical Association*, 295(4):429–433, 2006. doi: 10.1001/jama.295.4.429. URL `http://jama.ama-assn.org/content/295/4/429.abstract`.

George W. Cobb and Yung-Pin Chen. An application of Markov chain monte carlo to community ecology. *The American Mathematical Monthly*, 110(4):pp. 265–288, 2003. ISSN 00029890. URL `http://www.jstor.org/stable/3647877`.

William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *IIWeb*, pages 73–78, 2003.

Thomas E. Finucane and Chad E. Boult. Association of funding and findings of pharmaceutical research at a meeting of a medical professional society. *The American Journal of Medicine*, 117(11):842 – 845, 2004. ISSN 0002-9343. doi: 10.1016/j.amjmed.2004.05.029. URL `http://www.sciencedirect.com/science/article/pii/S0002934304005832`.

Lawrence M. Friedman, Curt D. Furberg, and David L. DeMets. *Fundamentals of Clinical Trials*, chapter 1, page 2. Springer New York, 4 edition, 2010.

Hyunmo Kang, Lise Getoor, Ben Shneiderman, Mustafa Bilgic, and Louis Licamele. Interactive entity resolution in relational data: A visual analytic tool and its evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 14(5):999–1014, 2008.

Maurice G. Kendall. *Rank Correlation Methods*. Charles Birchall & Sons ltd, London, 3 edition, 1962.

David A. Kessler, Janet L. Rose, Robert J. Temple, Renie Schapiro, and Joseph P. Griffin. Therapeutic-class wars – drug promotion in a competitive marketplace. *New England Journal of Medicine*, 331(20):1350–1353, 1994. doi: 10.1056/NEJM199411173312007. URL http://www.nejm.org/doi/full/10.1056/NEJM199411173312007.

Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, An Chi Guo, and David S. Wishart. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, 39(suppl 1):D1035–D1041, 2011. doi: 10.1093/nar/gkq1126. URL http://nar.oxfordjournals.org/content/39/suppl_1/D1035.abstract.

R. Luce and Albert Perry. A method of matrix analysis of group structure. *Psychometrika*, 14:95–116, 1949. ISSN 0033-3123. URL http://dx.doi.org/10.1007/BF02289146. 10.1007/BF02289146.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. doi: 10.1126/science.298.5594.824. URL http://www.sciencemag.org/content/298/5594/824.abstract.

M. E. J. Newman. The structure and function of complex networks. *SIAM REVIEW*, 45:167–256, 2003.

Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and János Kertész. Generalizations of the clustering coefficient to weighted complex networks. *Phys. Rev. E*, 75:027105, Feb 2007. doi: 10.1103/PhysRevE.75.027105. URL http://link.aps.org/doi/10.1103/PhysRevE.75.027105.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998. doi: 10.1038/30918.

Laszlo Zahoranszky, Gyula Katona, Peter Hari, Andras Malnasi-Csizmadia, Katharina Zweig, and Gergely Zahoranszky-Kohalmi. Breaking the hierarchy - a new cluster selection mechanism for hierarchical clustering methods. *Algorithms for Molecular Biology*, 4(1):12, 2009. ISSN 1748-7188. doi: 10.1186/1748-7188-4-12. URL http://www.almob.org/content/4/1/12.

Katharina Zweig and Michael Kaufmann. A systematic approach to the one-mode projection of bipartite graphs. *Social Network Analysis and Mining*, 1:187–218, 2011. ISSN 1869-5450. URL http://dx.doi.org/10.1007/s13278-011-0021-0. 10.1007/s13278-011-0021-0.

Katharina A. Zweig. How to forget the second side of the story: A new method for the one-mode projection of bipartite graphs. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '10, pages 200–207, Washington, DC, USA, 2010. IEEE Computer Society.

ISBN 978-0-7695-4138-9. doi: http://dx.doi.org/10.1109/ASONAM.2010.24. URL http://dx.doi.org/10.1109/ASONAM.2010.24.

A. Patterns for Intervention to DrugBank mapping

- ' and '
- ' then '
- ' plus '
- ' with '
- ' or '
- ' vs. '
- ' versus '
- '+'
- ','
- '/'
- '(+)
- '&'
- '('
- ')'
- '['
- ']'

Noise patterns:

- "®",
- "™",
- "©",
- "(P|p)lacebo"
- "\d{1,4} ?mc?(g|l)"

The last pattern is used to remove dosage information, such as “400 mg”

Source	Combination	Tokenizer	Synonyms	Precision	Recall	F-Score
DBPedia + DrugBank	Intersection	Default	No	0.9888	0.4251	0.5946
DBPedia	-	Default	No	0.9785	0.4396	0.6067
DrugBank	-	Default	No	0.9533	0.4928	0.6497
DBPedia + DrugBank	Union	Default	No	0.9459	0.5072	0.6604
DBPedia + DrugBank	Intersection	Default	Yes	0.9640	0.5169	0.6730
DBPedia	-	Default	Yes	0.9483	0.5314	0.6811
DrugBank	-	Default	Yes	0.9343	0.6184	0.7442
DBPedia + DrugBank	Union	Default	Yes	0.9225	0.6329	0.7507
DBPedia + DrugBank	Intersection	Space	No	0.9728	0.6908	0.8079
DBPedia	-	Space	No	0.9733	0.7053	0.8179
DrugBank	-	Space	No	0.9000	0.7826	0.8372
DBPedia + DrugBank	Union	Space	No	0.9016	0.7971	0.8462
DBPedia + DrugBank	Intersection	Space	Yes	0.9576	0.7633	0.8495
DBPedia	-	Space	Yes	0.9524	0.7729	0.8533
DrugBank	-	Space	Yes	0.8768	0.8599	0.8683
DBPedia + DrugBank	Union	Space	Yes	0.8738	0.8696	0.8717

Table 6: Results DrugBank ID look-up evaluation, sorted by F-Score in ascending order

B. Full evaluation results for drug linking

See table B for full evaluation results, including synonyms.

As an addendum to 6.2.1:

It is apparent that incorporation of intervention synonyms improves the F-Score considerably over configurations where only the intervention string itself was used.

C. Patterns for DrugBank to EDGAR mapping

- "inc"
- "co"
- "pharmaceuticals"
- "pharmaceutical"
- "ltd"
- "pharma"
- "farmaceutica"
- "sa"
- "products"

- "holdings"
- "holding"
- "div"
- "corp"
- "plc"
- "chemicals"
- "consumer"
- "healthcare"
- "the"
- "llc"
- "S.a.r.l."
- "Luxembourg"
- "Intermediate"
- "corp"
- "DE"
- "labs"
- "usa"
- "corporation"
- "sub"
- "and"
- "laboratories"
- "therapeutics"
- "oncology"
- "biotherapeutics"
- "products"
- "lp"
- "finance"

- "&"
- "capital"
- “company"
- "US"