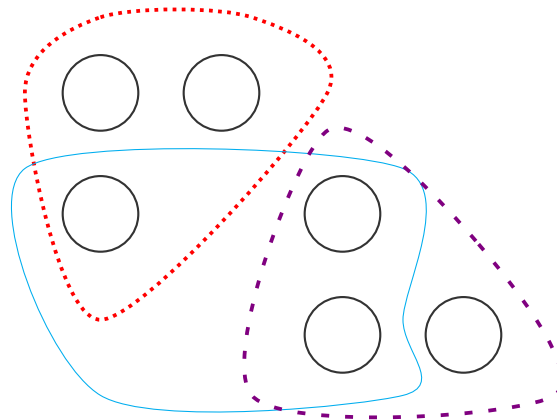


# Comparing Topological Communities and Communities of Interest Using Topic Modeling

Master Thesis  
Technical Artificial Intelligence



*Author:*  
Vincent TUNRU

*Supervisors:*  
Dr. Paola MONACHESI  
Dr. Ad FEELDERS

August 27, 2014



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Social Network Analysis . . . . .	1
1.2	Communities of Interest . . . . .	2
1.3	Research Question . . . . .	3
<b>2</b>	<b>Methodology</b>	<b>7</b>
2.1	Latent Dirichlet Allocation . . . . .	7
2.1.1	LDA extensions . . . . .	11
2.1.2	Relevant work on LDA and community detection . . . . .	14
2.2	Clustering in Social Graphs . . . . .	15
2.3	Cluster Validation . . . . .	21
2.3.1	Cluster structure . . . . .	21
2.3.2	Normalised Mutual Information . . . . .	23
2.3.3	Jaccard Similarity Coefficient . . . . .	24
<b>3</b>	<b>Detecting Communities of Interest</b>	<b>27</b>
3.1	reddit . . . . .	27
3.2	The gold standard . . . . .	31
3.3	Detecting communities . . . . .	34
3.4	Conclusions . . . . .	38
<b>4</b>	<b>Comparison to topological communities</b>	<b>39</b>
4.1	Enron . . . . .	39
4.2	The topological network . . . . .	40
4.3	Communities of interest . . . . .	42
4.4	Conclusions . . . . .	45
<b>5</b>	<b>Conclusions and Future Work</b>	<b>47</b>
5.1	Improve the gold standard . . . . .	48
5.2	Improve the quality of the data sets . . . . .	49
5.3	Incorporate LDA research . . . . .	50
5.4	Incorporate other features . . . . .	51



# Chapter 1

## Introduction

### 1.1 Social Network Analysis

Humans are social animals. During our lifetime, we meet and connect with many different people. By drawing people and their relationships as nodes connected through edges in graphs, social networks can be visualised. The study of these social networks is called *Social Network Analysis*, allowing the application of graph theory to social relationships (Otte and Rousseau, 2002). The type of relationships used can vary (Pineiro, 2011), from e.g. friendship to organisational relations.

Although there are many interesting characteristics of social networks to study, of particular interest to this thesis is the study of *community structure* within these social networks. Girvan and Newman (2002) provide the following definition of communities in graph-theoretical terms:

subsets of vertices within which vertex-vertex connections are dense,  
but between which connections are less dense.

In other words, communities are considered to be clusters of nodes in the graph that share many relations with other nodes in the same clusters. The type of communities described by this definition will be referred to as *topological communities* in this thesis. Intuitively, the definition makes sense. For example, consider the inhabitants of London. Many South Indians migrants in London live in the district of East Ham<sup>1</sup>. One would expect a proper definition of communities to classify these people as a community. By the definition above, one could draw a graph of the inhabitants of London by drawing edges between two persons indicating, for example, that they have met. It is to be expected that the South Indian migrants of East Ham have relatively often met. In other words: they would form a densely connected subgraph indicating, indeed, a community.

Although Girvan and Newman (2002) posit that these communities “might represent real social groupings, perhaps by interest or background”, this postulate has not yet been backed by substantial evidence. Thus, the graph-theoretical approach to finding community structure fails to properly explain

---

<sup>1</sup>See <https://neighbourhood.statistics.gov.uk>, retrieved on August 15th, 2014.

how to interpret these structures; what does being a member of a community mean?

## 1.2 Communities of Interest

With the advent of the internet, communities are no longer restricted geographically. People can connect from all over the world, realising *virtual communities* (Rheingold, 1993). As an example, on the social networking website Twitter, users can get engaged in political discussions (Conover et al., 2011) with people of whom they know nothing more than the username they use on Twitter. However, if these communities are not geographically bound, then the question remains what drives people to cluster together. In the field of sociology, many types of community have been proposed. One particular type of community is the *community of interest*. Henri and Pudelko (2003) define this as:

a gathering of people assembled around a topic of common interest. Its members take part in the community to exchange information, to obtain answers to personal questions or problems, to improve their understanding of a subject, to share common passions or to play.

The question arises whether the communities found by inspecting network topology as done in social network analysis conforms to any particular type of community. For example, Zhou et al. (2006) already wondered whether topology-based community discovery methods “suffer from the lack of semantic interpretation”, and that e.g. “given a group of email users discovered as a community, a natural question is why these users form a community?” There is an implicit assumption in some research that they are, in fact, communities of interest. This thesis is focused on finding out whether it is indeed the case that these topological communities bear some semblance to the communities of interest as defined above.

The expectation is that they will. In her seminal PhD thesis, Boyd (2002) argued that members of virtual communities manage sophisticated forms of identity management. People choose to display different parts of their identities in different contexts. This implies that, in different social groups, people’s use of language also adapts to the group. This is supported by Danescu-Niculescu-Mizil et al. (2013), who found that as someone “spends more time with the community, they adopt and start using the specific language of the community.” Some topics might be more appropriate to discuss with one group, while others might be more appropriate for other groups. This could be a result of different groups focusing on different interests. If topological communities represent these different social groups, and these social groups are centered around certain interests, then those social groups might also be what is understood as a community of interest.

Another reason to expect that topological communities and communities of interest might be similar is the research of Backstrom et al. (2006). The authors were interested in finding out what motivated people to join specific communities of interest. They found out that one is more likely to join a community when many of one’s friends are already a member of this community. In other words: members of a community of interest are likely to reel in people related

to them, shaping topological communities after communities of interest. A possible explanation could be that one of the qualities that makes a friend a friend, is the sharing of one's interests – making them more likely to be interested in the same communities. For example, a student of Artificial Intelligence could be a member of a community discussing news on the subject of artificial intelligence. This student is likely to have friends many of whom are also students of Artificial Intelligence, and thus are likely to share the interest. Their presence in the respective community hence correlates with the student's presence there.

Gathering insight into the interests of particular communities also has many practical applications. For example, Zhang et al. (2012) suggested that being able to accurately target specific groups of people can be very beneficial for viral marketing. For example, being able to promote a newly released game directly to gaming aficionados allows you to more easily connect those interested in buying it to those interested in selling it. Another case where many practical uses lie is in community management. As claimed by Kozinets (1999), “the more marketers can provide virtual community of consumption members with the meaning, connection, inspiration, aspiration, and even mystery and sense of purpose that is related to their shared consumption identities, the more those consumers will become and remain loyal.” Clear insight in the focus of communities and of which persons would share most interests together might allow for the enhancement of these persons' sense of purpose, thus strengthening member retention.

### 1.3 Research Question

The aim of this thesis is to find out whether the traditional topological communities can be found to be alike to what are defined as communities of interest in terms of classifying the same people as members of the same communities. This is guided by the following research question:

Do communities of interest align with topological communities?

There are two possible types of answers to this question, either of which provides valuable insights. The first possibility is that the two types of community are not alike. It is often assumed that topological communities share a common interest. That would then be proven to be inaccurate, which should be kept in mind when investigating either type of community in the future. No longer should topological communities then be interpreted to provide information about the interests of groups of people. On the other hand, if communities of interests were recognised as differing from topological communities, it could lead to them being researched on their own.

The other possible outcome is that the two types of communities are, in fact, similar. This would pave the way for better interpretation of the results of topological community-finding algorithms, allowing to use existing research on communities of interest to supplement that on topological community-finding algorithms. This could have benefits like being able to use topological community detection algorithms for researching communities of interests, when only information about user relationships is available.

The hypothesis is that these two types of communities do not align. A topological community, when considered on a graph of people, might very well

indicate a group of people of which most members know other members and regularly interact with them. However, the relationships between these people need not necessarily indicate a common interest. For example, in a graph of people where the edges indicate that two persons have spoken to each other, members of the same family are likely to be found to be a community based on the graph's topology. However, the interests of family members could greatly diverge.

Finding the answer to this question is done in two steps, each with its own sub-question. The first sub-question is:

Can communities of interest be detected by interpreting textual content?

The application of an existing algorithm to the discovery of communities of interest is investigated. Even if the outcome of this research is that communities of interest are distinct from topological communities, being able to identify CoIs is still useful. The re-use of an existing algorithm for this new purpose also allows for the re-use of existing research to improve it.

The algorithm used is a *topic modeling* algorithm: a way of classifying text into separate topics based on their content. It will be explained elaborately in Section 2.1. The expectation is that such an algorithm will be able to detect communities of interest. As members of the same community of interest are expected to have the respective interest as the main discussion topic, an algorithm for finding topics should be able to find these communities.

The second step is comparing the results of the above algorithm to a traditional algorithm for topological community detection. This is driven by the following sub-question:

Can topological communities be detected by interpreting textual content?

If it turns out that the two algorithms produce similar results, then the two types of communities can be said to be similar as well. This means first of all that existing assumptions about their being equivalent would be validated. Furthermore, the algorithms would be able to complement each other by providing information about both the graph structure of a community and its interests.

However, when the results deviate, they would give reason to question the assumption of the equality of the two types of community. There would be reason to clearly differentiate between the two types of community in future research, and the two algorithms could be more consciously used in different applications. It would also provide additional reason to further investigate algorithms for detecting communities of interest and their respective interests, an area of research that currently receives only little attention compared to that of detecting topological communities.

As said, the hypothesis is that the two types of communities have different compositions. Since the expectation is that communities of interested will be detected from textual content, the expectation follows that topological communities will not.

This thesis is structured as follows. It starts with a discussion of the methodology in Chapter 2, which provides an overview of the algorithms used in this thesis and discusses the appropriate relevant literature.



Then, Chapters 3 and 4 discuss the first and second sub-questions, respectively, in three parts. First, they provide an overview of the data set used for the specific experiments to answer each sub-question. This is followed by a brief overview of the algorithms used in that experiment, explaining how they are used and what their output looks like. Then finally, the results are evaluated and conclusions are drawn.

These chapters are followed by an evaluation of their results, and using them to draw conclusions about the main research question. The thesis end with suggestions for future research based on the research done for this thesis.



## Chapter 2

# Methodology

This chapter introduces the algorithms used in this thesis. It includes short explanations of the algorithms and what they can be used for, and presents relevant literature on the subject. The first algorithm discussed is Latent Dirichlet Allocation (LDA), a topic modeling algorithm that will be used to detect communities of interest. It is followed by a brief summary of related algorithms and to what extent they are relevant to the work in this thesis. Finally, an overview is provided of how the algorithm can be used to detect communities.

The section on LDA is followed by Section 2.2 on Girvan and Newman’s community discovery algorithm, a popular method of finding topological communities. This is the algorithm that will be used as a benchmark to compare the application of LDA to. This is followed by a description of the notion of modularity, a concept to measure how tightly knit communities in graphs are. Since the algorithm results in several possible community clusterings, this measure is often used to select the most convincing one.

Finally, two methods of cluster validation will be discussed. These methods are needed to evaluate the predictive value of the algorithms. In the case of LDA, they make it possible to assess how well its detected clustering conforms to the actual clustering. For the Girvan-Newman algorithm they allow for comparing the resulting clustering to LDA’s.

### 2.1 Latent Dirichlet Allocation

*Topic modeling* is a form of unsupervised machine learning most often used to discover “topics” in sets of documents. The basic intuition behind topic modeling is that a topic is associated more with certain words than with others. For example, the topic “pets” brings to mind words like “cat” and “dog”, but not “college” or “administration”. Likewise, documents concerning pets and health are expected to contain a relatively high percentage of words strongly associated with those topics. The assumption in topic modeling is that, if one observes words commonly occurring together in large corpora, then one should be able to reconstruct reasonably well the original topics associated with those words.

Arguably the most commonly used topic model is **Latent Dirichlet Allocation** (LDA), pioneered by Blei et al. (2003). At the basis of LDA is a

*generative model*, a hypothetical description of how a corpus of documents is created. The biggest simplification made by the model is that it is a *bag of words model*: it assumes that the order of the words in a document does not matter; they are simply “bags” of words. Although this makes the model inaccurate in terms of describing the actual process of writing documents, it would be possible for a human to view the documents generated by the model and to coarsely guess the topics it discusses, despite the shuffled word order. A useful aspect of LDA being a generative model is that it can not only be used to classify existing documents in a corpus, but can also be used to categorise new, formerly unseen documents added to the corpus.

LDA’s generative model is a process that works as follows. First of all, the model posits the following elements:

1. A vocabulary: a list of words that can be used to write the documents in the corpus.
2. A list of topics, each topic being a probability distribution over the words in the vocabulary indicating the likeliness of that word to be written when talking about that topic.
3. For each document that is to be generated, a probability distribution over the topics that determine to what extent that document is about each topic.

Then, for each document that is to be generated, the following process is assumed to occur:

1. For each word token that is to be generated:
2. Pick a topic from the probability distribution over topics for this document.
3. Pick a word from the probability distribution over the vocabulary for this topic.
4. Repeat steps 1-3 until all the words of this document have been generated.

An example of how a document would have been generated is shown in Figure 2.1, courtesy of Blei (2012). On the left-hand side there is a list of topics (each in its own colour), showing the probabilities associated with certain words for those topics. On the right-hand side there is a document with an associated distribution over topics. Each word token in the document was generated by choosing one of the topics according to that distribution, and then picking the word from the distribution associated with that topic to be used for that word token.

The initial distributions over words (for topics) and over topics (for documents) are taken from Dirichlet prior distributions parameterised by  $\beta$  and  $\alpha$ , respectively. The Dirichlet distribution is the conjugate prior of the multinomial distribution, making it a suitable choice for providing prior probabilities for words in topics and topics in documents.  $\beta$  and  $\alpha$  are vectors containing the concentration parameters for each word and topic, respectively, determining how evenly distributed the resulting distribution is. In this case, since no prior information is available on what words/topics will be more prominent, every

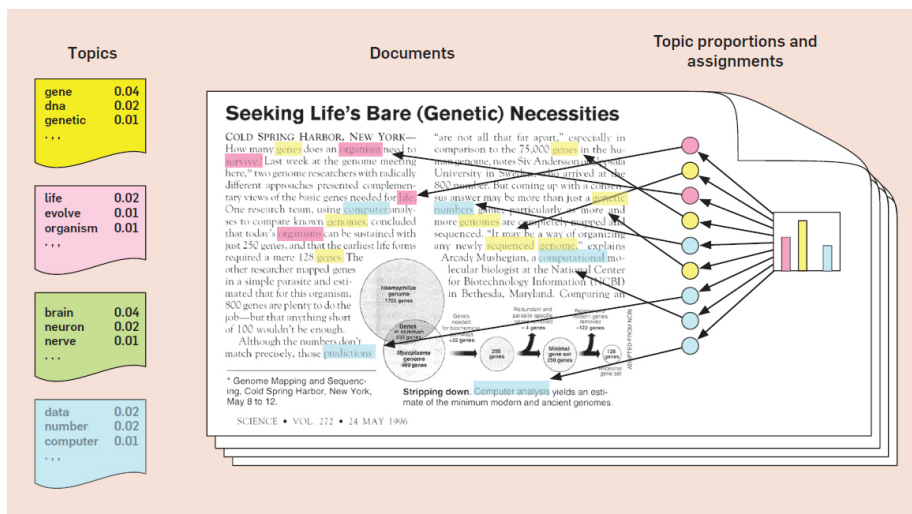


Figure 2.1: An example of a document as generated by the model. Topics assign probabilities to words, for each word token in a document a different topic is selected according to that document’s probability distribution over topics, from which the respective probabilities are used to determine which word to use. Example by Blei (2012).

element in each vector has the same value, resulting in a symmetric Dirichlet distribution. A low value ( $\alpha \ll 1$  or  $\beta \ll 1$ ) means that it is likely to result in probability distributions that have most of their mass concentrated on one of their components. This is desired in the case of LDA, where topics are characterised most strongly by only a small percentage of the words in the vocabulary, and documents discuss only a few of the possible topics.

Figure 2.2 shows the graphical model (using plate notation) that LDA assumes to apply to the generative process of documents in a corpus. Here, the top plate represents the topics:  $T$  distributions  $\phi$  drawn from a Dirichlet distribution. Below that there are two nested plates. The outer plate represents the  $M$  documents, and contains their associated distributions  $\theta$  over the topics, also drawn from a Dirichlet distribution. The inner plate represents the  $N$  word tokens. For each word, topic  $z$  is selected from the document’s distribution  $\theta$ . This topic has an associated distribution over words  $\phi_z$  that determines which word is used as word token  $w$ .

LDA tries to fit this model to the data to uncover the hidden variables based on the only observed variable (shaded): the actual words in the documents of the training set. Primarily, it tries to fit the composition of the topics (i.e. the distributions over the vocabulary) and, for each word in each document, the topic that generated it. A commonly used method – and the one implemented in MALLET, the software package used for this thesis – is Gibbs sampling.

Using Gibbs sampling to estimate the topics was first proposed by Griffiths and Steyvers (2004). The parameters that are to be estimated are the topics’ probability distributions over words ( $\phi$ ), and the documents’ probability distributions over topics ( $\theta$ ). These parameters are approximated from estimates of the assignments of words to topics. Gibbs sampling generates a number of

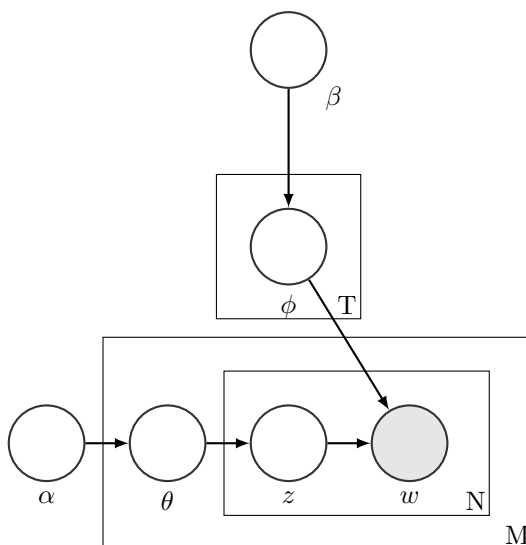


Figure 2.2: Graphical model of LDA.

samples that, after a burn-in period, converges towards the corpus' assumed distribution. Each sample fits a new topic assignment for each word token in the corpus as follows.

Considering a word token  $i$ , the conditional distribution for that word to be assigned to topic  $j$  is  $P(z_i = j | \mathbf{z}_{-i}, w_i, d_i, \cdot)$ , with  $w_i$  being the specific word that is used for word token  $i$ , and  $d_i$  the document in which this word is contained.  $\mathbf{z}_{-i}$  contains the topic assignments for all other word tokens, and  $\cdot$  all other known/observed information. It can be calculated from the word count matrices as shown in equation 2.1. The matrix  $C^{WT}$  of size  $W \times T$  (vocabulary size times the number of topics) tallies the number of word tokens containing specific words that are assigned to specific topics, and  $C^{DT}$  of size  $D \times T$  (number of documents times the number of topics) tallies the number of word tokens in each document that are assigned to specific topics.

$$P(z_i = j | \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^W C_{w j}^{WT} + W\beta} \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^T C_{d_i t}^{DT} + T\alpha} \quad (2.1)$$

Here,  $C_{w j}^{WT}$  is the number of word tokens, excluding  $i$ , that are word  $w$  and are assigned to topic  $j$ . Likewise,  $C_{d_i t}^{DT}$  are the number of word tokens, excluding  $i$ , that are in document  $d_i$  and are assigned to topic  $t$ . This value is calculated for every possible topic, and each value is divided by the sum of all the values of all possible topics for this word token to get the actual probability that the topic for  $i$  is  $j$ .

The equation can be understood as being composed of two parts: the probability of the word for word token  $i$  being  $w_i$  given a topic  $j$ , and the probability of the topic being  $j$  given that the word token is in document  $d_i$ . The consequences of this is that, if this word is more often labeled as being from topic  $j$ , this word has a higher probability of being assigned to that topic as well. Like-

wise, if words in document  $d_i$  are often assigned to topic  $j$ , then that increases the probability of this word being assigned to that topic as well.

Initially, all word tokens are randomly assigned to a topic. Then, each sample recalculates the assigned topic for each word token based on the other words' assignments as described above. After the burn-in period, estimates  $\phi'$  and  $\theta'$  of the word-topic distributions and the topic-document distributions, respectively, can be obtained from the counts of a sample as shown in equation 2.2.

$$\phi'_i{}^{(j)} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^W C_{kj}^{WT} + W\beta} \quad \theta'_j{}^{(d)} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^T C_{dk}^{DT} + T\alpha} \quad (2.2)$$

They are the distributions of the predictions that a new token of word  $i$  would be classified as being from topic  $j$  (for  $\phi'_i{}^{(j)}$ ), and that a new token in document  $d$  would be from topic  $j$  (for  $\theta'_j{}^{(d)}$ ).

Griffiths and Steyvers (2006) provided the example from Figure 2.3. It shows the topic assignments of the words in a small corpus in two samples: one at the start of the Gibbs sampling procedure (at the top) and one after 64 iterations. Each circle represents a word token; the actual word is given by the column it is in, the document it is in is given by the row, and the topic it is classified as in that sample is determined by whether it is shaded or not (in this example, there are only two topics). Initially, each word is randomly classified in either of the topics. As the words *river* and *stream* commonly occur in the same documents together, they converge towards being classified as being about the same topic; the same goes for *money* and *loan*. The word *bank* co-occurs with words from both of the topics; since the word itself is ambiguous, its classification is mostly determined by the topics dominant in the relevant document.

In this thesis, LDA is used for community detection. It is a fairly straightforward choice, as the basic assumptions about the generation of documents transfer well to the generation of text by people. Furthermore, it is a widely used and researched algorithm, with several libraries available that make working with it easy. There is also much research on LDA and ways to improve it that could be utilised in future research. For the sake of consistency with other literature, the chapter uses the topic-document terminology. It should be noted, though, that the generative model has to be altered slightly to understand its use in this thesis. Specifically, one has to replace “topic” with “community” and “document” with “person” in the generative model. This means that instead of on the generation of documents, the process is focused on the generation of text by the modeled user.

### 2.1.1 LDA extensions

One of the main advantages of LDA is that, as a probabilistic model, it is easily extensible. An overview of many of these extensions was compiled by Blei (2012). Some of these extensions focus on relaxing and extending the assumptions made by LDA in an attempt to discover more sophisticated structures in text. For example, the bag-of-words-assumption by LDA is fine when the goal is to uncover the general topical structure of a text, but is woefully inadequate when the order of words matters, such as when generating language. Another

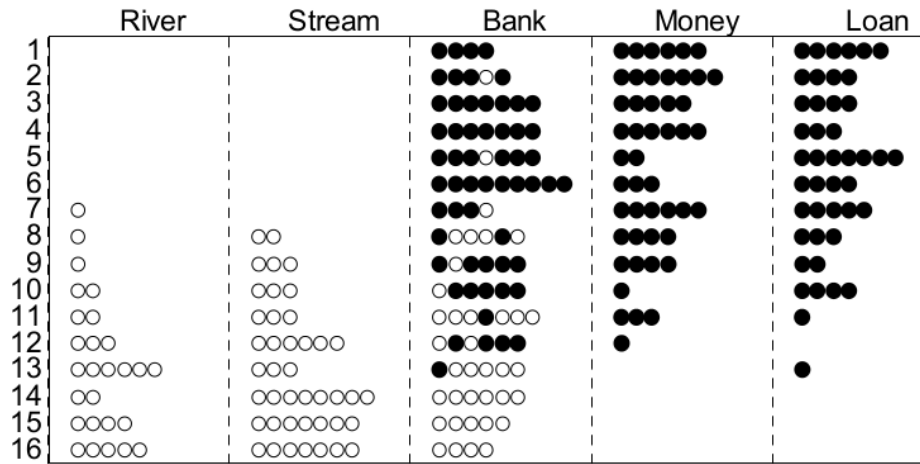
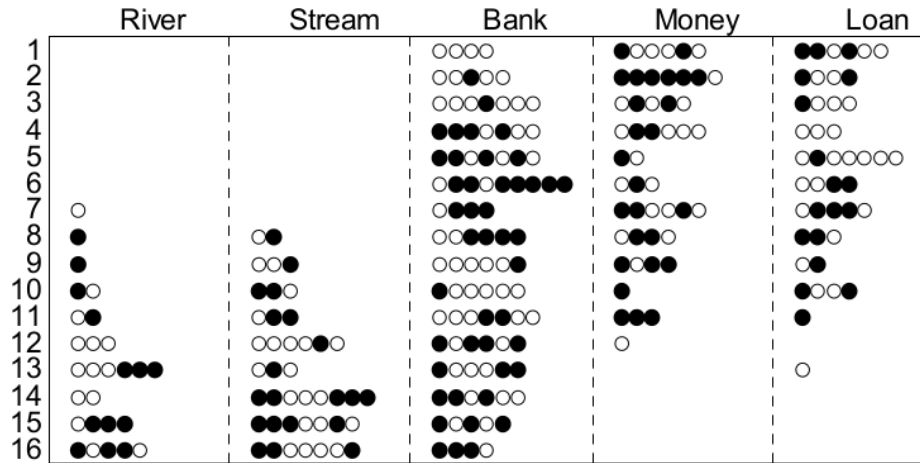


Figure 2.3: An example of the topic assignments of two samples of the Gibbs sampling procedure, by Griffiths and Steyvers (2006).



example of how LDA can be extended by relaxing its assumptions is the *Correlated Topic Model* by Lafferty and Blei (2006). This extension allows for topics to be related – a document about computer science would be more likely to also be about mathematics than about sports. Although it would be interesting to investigate of what use many of these extensions could be when discovering communities of interest, investigating them unfortunately is outside the scope of this thesis.

Other extensions focus on incorporating other data into the model. For example, besides the document words appear in, many data sets also include other information, such as the date the document was written or published. This data might provide additional hints about the topics of the documents, and thus incorporating the data into the model could improve classification accuracy. It also works the other way around: the model could be used to make assertions about the added data. To follow up on our example, it could be inferred which topics were often discussed in certain time periods, but not in others. Two of these extensions that incorporate additional data deserve some closer inspection, because they seem at first sight to be relevant to this thesis. This section discusses them briefly, and explains why it was decided not to use them in this thesis.

The first of these is the *author-topic model* from Rosen-Zvi et al. (2004). This model extends LDA by incorporating not only the documents words are from, but also the person that wrote them. This allows for, for example, the computation of the similarity of different authors in terms of the topics they write about. It results in the updated graphical model shown in Figure 2.4. Compared to plain LDA, the author-topic model alters the graphical model through the addition of a few extra variables. No longer are the topics in a document chosen directly from a single topic-document distribution. Instead, there is now a separate plate for the author-specific topic-document distributions. The document’s author  $\mathbf{a}_d$  determines distribution  $\theta_x$  that will be used for picking the document’s topic index  $z$ .

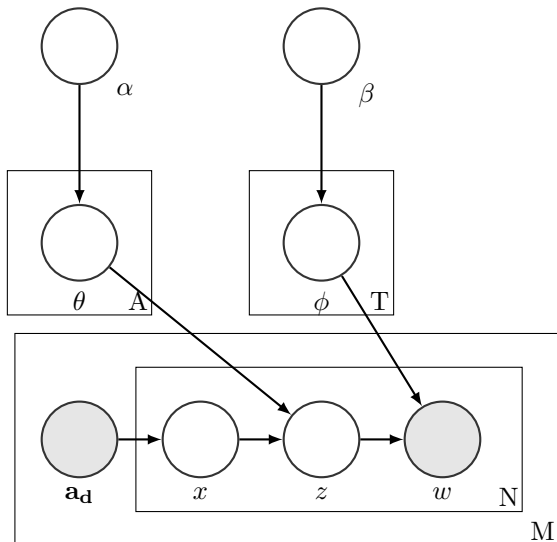


Figure 2.4: Graphical model of the author-topic model.

The *community-author-topic model*<sup>1</sup> (CAT), by Mimno et al. (2007), is an extension of the author-topic model and, hence, of LDA. However, instead of adding another observed variable such as the document’s author, the CAT model assumes an additional latent variable to be present. In addition to words being part of documents, and documents being written by authors, these authors are also claimed to be part of a community. This, the authors assume, influences the topics the documents are about as well. In other words: these “communities” could, perhaps, be seen as communities of interests.

So why, then, does this thesis make use of plain LDA? The reason is that this thesis focuses on classifying persons in communities of interest. In other words, there is no interest in individual documents, and communities are defined by the topics they discuss. So considering that LDA is commonly used as a *topic* model for classifying *documents*, and that the *author-topic* model and the *community-author-topic* model classify *documents* as well, LDA in this thesis can be seen as being used as a *community* model for classifying *authors*. The above two models, then, add the classification of documents and topics, making them superfluous.

## 2.1.2 Relevant work on LDA and community detection

Apart from the extensions to LDA that incorporate community structure into the classification of documents, there is also work by Zhang et al. (2007) that directly uses LDA for finding community structure instead of classifying documents. Its approach fundamentally differs from that used in this thesis, however. It does not look at text produced by people, but considers the occurrence of *social interactions*, or specifically: whether two people co-authored a paper together. A list of interactions by an author were considered the “documents”, with the latent topic distribution representing the communities – just as in this thesis. This means that, although it uses LDA, this work still essentially focuses solely on topological communities instead of communities of interest. To find communities, it incorporates social interactions (the relationships), but discards semantic information (the content of the papers).

There does not appear to be much research that uses topic modeling to actually discover community structure from semantic information. However, there is some research that uses it to complement traditional methods for discovering topological community structure with semantic information. As an example, Li et al. (2010) applied LDA after applying topological community detection methods to see how topically consistent topological communities are over time. They found out that the top words in the topics used by authors in a community over time remain consistent, and that the topics by the authors from different communities differed as well.

The only research found that makes a distinction between topological communities and communities of interest (they refer to them as “topical communities”) are Ding (2011); Ghali et al. (2012). In the relevant experiments, the detection of “topical communities” is done based on the author-topic model. They take as communities authors that commonly discuss the same topics, while discarding the document classification. In principle this is similar to the method of detecting communities of interest in this research. Unfortunately, the question

---

<sup>1</sup>The  $CUT_1$  and  $CUT_2$  models from Zhou et al. (2006) are based on similar ideas. Since the same reasoning for not using it applies as for the community-author-topic model, they are not discussed separately.

of whether the two align is left open, nor are the two types of communities compared directly. Instead, the research focuses on how well different topics can be discerned within their detected topological communities, and whether different topological communities can be detected within communities of interest.

Thus, the application of topic modelling to the discovery of community structure is a novel one. Nonetheless, the few modifications it requires to the generative model of LDA are intuitively reasonable and allow for the straightforward application of LDA to this new domain, and shows a lot of promise to be able to deliver good results.

## 2.2 Clustering in Social Graphs

Whereas the previous section discussed LDA, the algorithm to be used to discover the composition of communities of interest, this section will deal with the method to be used for discovering community structure in graphs. There is a lot of research focusing on detecting such topological communities. Its aim is to find structures in social networks consisting of nodes of people connected through *ties* representing relationships between these individuals. Topological communities are often defined as sets of nodes on a social graph that are densely connected within the community, and sparsely connected to nodes outside the community. An example is shown in Figure 2.5. There are two communities: one of which the member nodes are unshaded, and one of which the member nodes are shaded. As can be seen, within each community there are many connections, whereas there are only few connections between nodes from different communities (these inter-community connections are indicated by dotted lines).

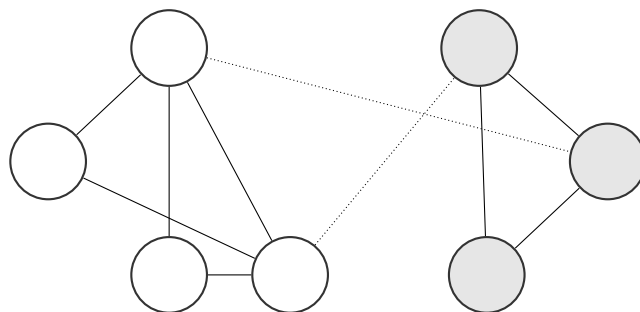


Figure 2.5: Example of two topological communities (shaded and unshaded).

One of the main methods of topological community detection is the work of Girvan and Newman (2002). Because this one is widely used and has readily available packages implementing the algorithm, this algorithm was chosen to be the benchmark for topological community detection. It will be used to answer the second sub-question of this research. The algorithm revolves around the concept of *edge betweenness*. The betweenness of an edge is defined as the number of shortest paths between different pairs of nodes in the (social) graph that run along that edge. Since nodes in different communities are often connected through edges between two nodes of different communities, the shortest paths are also likely to run via those edges. Hence, edges with a high edge betweenness are more likely to separate different communities from each other.

The Girvan-Newman algorithm takes advantage of this by gradually removing the edges with the highest edge betweenness, separating communities from each other into disconnected graphs. The steps of the algorithm are as follows<sup>2</sup>:

1. Calculate the betweenness of every edge in the network.
2. Remove the edge with the highest betweenness.
3. Recalculate the betweenness of the remaining edges.
4. Repeat the last two steps until all edges have been removed.

Figures 2.6 to 2.11 show a couple of iterations of the algorithm for a small social network, each iteration resulting in a potential community clustering with a different number of clusters.

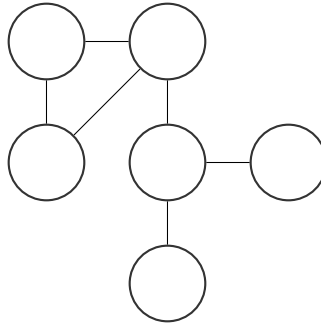


Figure 2.6: Girvan-Newman algorithm: In the first iteration, no edges have been removed yet – all nodes are in the same community.

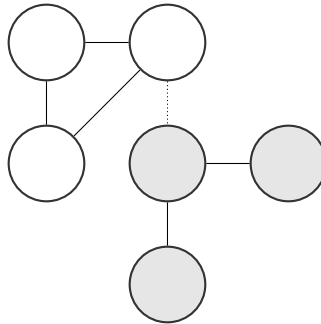


Figure 2.7: Girvan-Newman algorithm: In the second iteration, the edge is removed that is traveled along most often when considering all the paths between all pairs of nodes (marked by a dotted line). This separates the nodes in two clusters (communities), a shaded and an unshaded one.

Since several different potential community clusterings emerge as more edges are removed, a way is needed to determine which of these potential clusterings

<sup>2</sup>Algorithm 2.1 at the end of this Section provides an overview in pseudocode, including the selection criterion explained later in this Section.

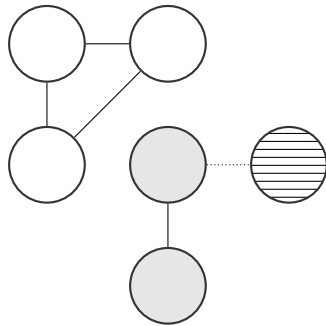


Figure 2.8: Girvan-Newman algorithm: The third iteration sees a node splitting off from the shaded community, forming its own singleton community.

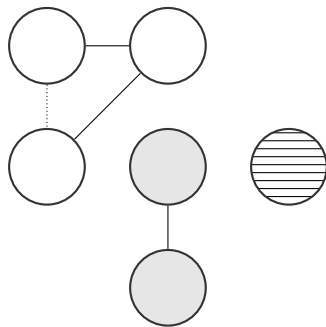


Figure 2.9: Girvan-Newman algorithm: In the fourth iteration, several edges have the highest betweenness. One of them is chosen arbitrarily and removed. Since this edge is not the only path through which its connected nodes can connect, it does not result in the creation of a new community.

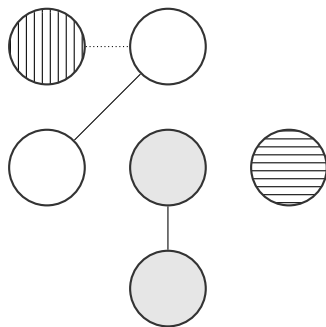


Figure 2.10: Girvan-Newman algorithm: In the fifth iteration, again, several edges have the highest betweenness. Removing one of them creates yet another community.

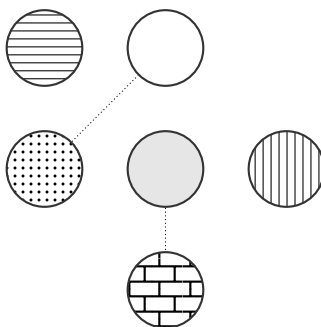


Figure 2.11: Girvan-Newman algorithm: In the sixth and seventh (final) iterations, removing the last two lines results in each node ending up being the only node in its community.

is best. After all, the clustering in the first iteration merely includes all nodes in a single community and is thus unsatisfactory, while the last iteration places each node in a different community, also providing only little information. A potential solution to this problem is mentioned in the followup paper (Newman and Girvan, 2004), where the authors describe the concept of *modularity*. It is based on whether the ratio of intra- vs. inter-community edges is higher than when they would have been assigned randomly. In other words: community assignments such that a large proportion of edges are between vertices from the same communities have a high score. By calculating the modularity of the resulting clustering after each edge removal, the best clustering can be selected.

To explain the concept of modularity, it is instructive to first realise that it tries to maximise the number of intra-community edges, and minimise the number of inter-community edges. Given a clustering, modularity is built around two variables:

1. The fraction of edges that link nodes in community  $i$  to nodes in community  $j$ :  $e_{ij}$ .
2. The fraction of edges that connect to at least one node in community  $i$ :  $a_i$ .

Now note that, if the community assignments are completely unrelated to the number of intra- and inter-community edges, the fraction of edges connecting nodes from community  $i$  to community  $j$  (so  $e_{ij}$ ) should approach  $a_i a_j$ , whereas this is not the case for better community clusterings. The modularity  $Q$  can then be defined as in equation 2.3 to satisfy the properties of having a higher value when there are many intra-community edges, and a lower value when the clustering gets closer to the expected value of a random clustering.

$$Q = \sum_i (e_{ii} - a_i^2) \quad (2.3)$$

To return to the example from Figures 2.6 to 2.11, Table 2.1 outlines the modularity of the clustering of each iteration. Note that although the modularity is based on the community assignments from each iteration, it considers edges as they are *in the unmodified network*, i.e. as they are at the start of the

first iteration. Let's consider the modularity of iteration 3 to illustrate the calculation. Figure 2.8 shows that this iteration resulted in a clustering with three communities, consisting of the white, shaded and striped nodes, respectively. Since the edges from the full network need to be considered, Figure 2.6 shows that there are six edges in total. Four of those edges link to at least one white node, and three of those are between two white nodes. This means that the white community's share in the modularity equation is  $\frac{3}{6} - \frac{4^2}{6}$ . Likewise, when  $i$  indicates the shaded community,  $e_{ii}$  is  $\frac{1}{6}$  (there is one edge between shaded nodes) and  $a_i$  is  $\frac{3}{6}$  (there are three edges connected to at least one shaded node). Since there is only a single striped node, there are no edges between two striped nodes. As there is only one edge connected to this one striped node, the striped community's part of the equation is  $\frac{0}{6} - \frac{1^2}{6}$ . Altogether this leads to the equation listed for the third iteration in Table 2.1.

As can be seen, the clustering with the highest (and, in fact, only positive) value of  $Q$ , namely  $\frac{5}{36}$ , is the one resulting from iteration 2. In other words: according to modularity, the best clustering is the one with two communities, with the community membership assignments denoted by the shading in Figure 2.7.

Algorithm 2.1 provides an overview of the Girvan-Newman algorithm, with modularity as a selection criterion, in pseudocode. As can be seen, the algorithm is fairly straightforward. It makes intuitive sense, is widely used, and has been proven by its authors to provide good results. All this makes it the algorithm of choice for use in this thesis for the discovery of topological communities.

---

**Algorithm 2.1** Pseudocode for the Girvan-Newman algorithm

---

```

removedEdges  $\leftarrow \emptyset$ 
repeat
  for all edges  $i$  do
    if  $i$  gives the highest EDGE.BETWEENNESS( $i$ ) then
      edgeToRemove  $\leftarrow i$ 
    end if
  end for
  removedEdges  $\leftarrow \text{CONCAT}(\textit{removedEdges}, \textit{edgeToRemove})$ 
  if removedEdges gives the highest MODULARITY(removedEdges) then
    edgesToRemove  $\leftarrow \textit{removedEdges}$ 
  end if
until number of edges = 0
   $\triangleright$  Removing these edges results in the desired clustering:
return edgesToRemove

```

---

Iteration	Modularity $Q$
1	$\frac{6}{6} - \frac{6^2}{6} = 0$
2	$\left(\frac{3}{6} - \frac{4^2}{6}\right) + \left(\frac{2}{6} - \frac{3^2}{6}\right) = \frac{5}{36}$
3	$\left(\frac{3}{6} - \frac{4^2}{6}\right) + \left(\frac{1}{6} - \frac{3^2}{6}\right) + \left(\frac{0}{6} - \frac{1^2}{6}\right) = -\frac{1}{18}$
4	$\left(\frac{3}{6} - \frac{4^2}{6}\right) + \left(\frac{1}{6} - \frac{3^2}{6}\right) + \left(\frac{0}{6} - \frac{1^2}{6}\right) = -\frac{1}{18}$
5	$\left(\frac{0}{6} - \frac{2^2}{6}\right) + \left(\frac{1}{6} - \frac{4^2}{6}\right) + \left(\frac{1}{6} - \frac{3^2}{6}\right) + \left(\frac{0}{6} - \frac{1^2}{6}\right) = -\frac{1}{2}$
6	$\left(\frac{0}{6} - \frac{2^2}{6}\right) + \left(\frac{0}{6} - \frac{2^2}{6}\right) + \left(\frac{0}{6} - \frac{2^2}{6}\right) + \left(\frac{1}{6} - \frac{3^2}{6}\right) + \left(\frac{0}{6} - \frac{1^2}{6}\right) = -\frac{4}{9}$
7	$\left(\frac{0}{6} - \frac{2^2}{6}\right) + \left(\frac{0}{6} - \frac{2^2}{6}\right) + \left(\frac{0}{6} - \frac{2^2}{6}\right) + \left(\frac{0}{6} - \frac{3^2}{6}\right) + \left(\frac{0}{6} - \frac{1^2}{6}\right) + \left(\frac{0}{6} - \frac{1^2}{6}\right) = -\frac{23}{36}$

Table 2.1: The modularity of each iteration of the example of Figures 2.6 to 2.11.



## 2.3 Cluster Validation

With the different tools for finding community structures at hand, what is still missing is a way to determine how well they do their job. Since finding communities among a set of people is basically just a method of clustering, cluster validation methods can be used to evaluate performance.

There are several characteristics of clustering methods that determine which validation tools might be applicable for each method. While for an elaborate overview one should refer to Manning et al. (2008), this section starts with a brief summary of the important aspects for this thesis. This is followed by an overview of the two validation methods that are used in this thesis: Normalised Mutual Information and the Jaccard Similarity Coefficient.

### 2.3.1 Cluster structure

Clusters can be hierarchical or flat. In a hierarchical clustering, clusters higher up in the hierarchy are composed of combinations of clusters lower down in the hierarchy. Figure 2.12 illustrates this: each level in the tree is a clustering, with each node in that level representing a single cluster (i.e. group of objects).

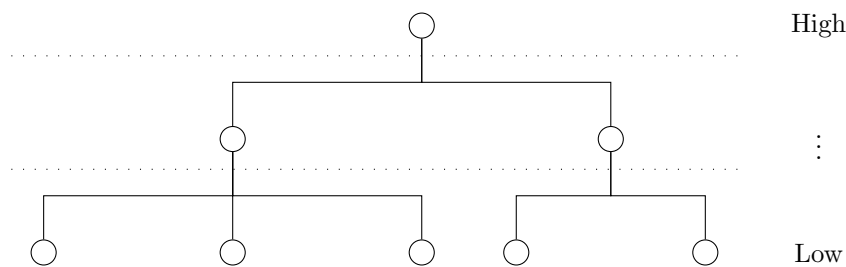


Figure 2.12: Example of hierarchical clusters, each level in the hierarchy representing a clustering. Note that the nodes here are clusters (i.e. a set of objects), not objects themselves.

Flat clustering methods, however, result in sets of clusters that have no structure to relate them together, and that make no claim about the relationship between objects. Clusters are not composed of other clusters, but only of individual objects. Essentially, each cluster is just a list of objects included in that cluster. Figure 2.13 shows an example of a flat clustering.

Apart from being hierarchical or flat, there is another property that differentiates clusters: they can either be hard or soft. A hard clustering means that each object is a member of exactly one cluster – no more, no less. For example, the objects in Figure 2.13 are part of one cluster or the other; the clusters are non-overlapping.

On the other hand, soft clustering means objects can be a member of multiple clusters, to varying extents. For example, a population could be clustered according to the languages they speak. People can speak multiple languages, and with varying skill. Figure 2.14 shows a clustering where several objects are in multiple clusters.

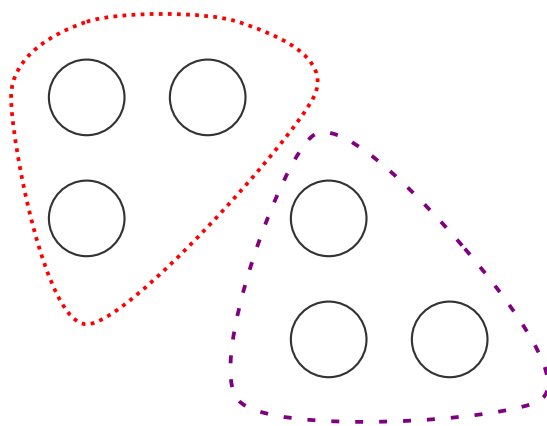


Figure 2.13: Example of a flat, hard clustering with two clusters.

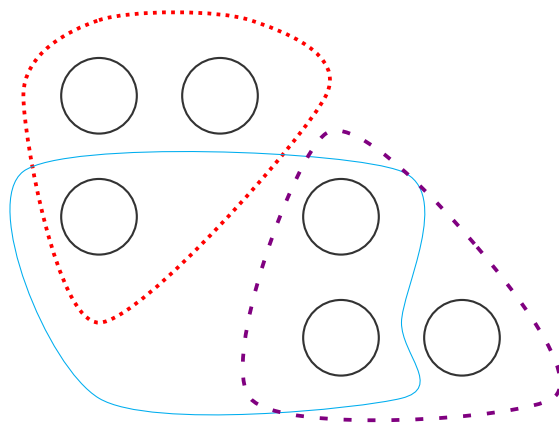


Figure 2.14: Example of a flat, soft clustering with three clusters, partially overlapping.

### 2.3.2 Normalised Mutual Information

The answer to sub-question one (“Can communities of interest be detected by interpreting textual content?”) will involve both community structure as discovered through LDA, and a real-life community clustering. Both of these are flat clusterings: there is no hierarchical structure among the clusters. Furthermore, people can be members of multiple clusters, to varying extents – i.e. the result is a soft clustering. A similarity measure to compare the two clusterings that can incorporate all this is Normalised Mutual Information (NMI). This measure revolves around the concept of how much information one clustering method can provide about the other. If two clusterings are the same, then you can predict the composition of one by looking at the other, since it is equal to it. If two clusterings are completely unrelated, then having one does not contribute to finding out the clustering of the other. Normalised Mutual Information takes two clusterings and returns a number between 0 and 1 to codify the amount of information they provide about each other – closer to 1 is better, with a value of 1 signifying perfect overlap. Hence, the method is perfect for comparing how well different clustering methods can detect actual clusterings. A disadvantage is that the results are not easily interpretable. Unfortunately, no methods were found that could evaluate soft clusterings *and* provide results that were meaningful on their own, without comparison to other methods.

Considering clusterings  $\Omega$  and  $\mathbb{C}$ , NMI is defined as shown in equation 2.4 (Manning et al., 2008):

$$NMI(\Omega, \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2} \quad (2.4)$$

Basically, the numerator is Mutual Information, and the denominator normalises this to a value between 0 and 1. Mutual information  $I$  is defined in equation 2.5, and entropy  $H$  is defined in equation 2.6.

$$I(\Omega; \mathbb{C}) = \sum_k \sum_j P(\omega_k \cap c_j) \log \frac{P(\omega_k \cap c_j)}{P(\omega_k)P(c_j)} \quad (2.5)$$

$$H(\Omega) = - \sum_k P(\omega_k) \log P(\omega_k) \quad (2.6)$$

Usually, the values in these equations can be calculated directly from membership counts:  $P(\omega_k)$  is simply the number of objects in cluster  $k$  divided by the total number of objects. This thesis, however, does not deal with probabilities of being a member of a certain cluster; the numbers represent an object’s involvement in a certain cluster, relative to the other clusters. To overcome this problem, the values are calculated using the following custom methods.  $P(\omega_k)$  is the mean of all objects’ ratio of being a member of cluster  $k$  in clustering  $\Omega$ . Likewise,  $P(c_j)$  is the mean of all objects’ ratio of being a member of cluster  $j$  in clustering  $\mathbb{C}$ . To calculate  $P(\omega_k \cap c_j)$ , for each object we multiply the odds of being a member of cluster  $k$  with those of being a member of cluster  $j$ . After calculating this for each object, we take the mean of all objects to be the odds of any object being a member of both clusters  $k$  and  $j$ , i.e.  $P(\omega_k \cap c_j)$ .

As an example, take the clustering in Table 2.2. To calculate  $P(\omega_1)$  we take the mean of all objects’ memberships of cluster 1 in clustering  $\Omega$ : about 4.33. Likewise, to calculate  $P(c_b)$ , we take the mean of all objects’ memberships of

Object #	Comm. 1	Comm. 2
1	0.2	0.8
2	0.4	0.6
3	0.7	0.3

(a) Clustering  $\Omega$ 

Object #	Comm. a	Comm. b
1	0.7	0.3
2	0.6	0.4
3	0.1	0.9

(b) Clustering  $\mathbb{C}$ 

Table 2.2: Example clustering of three objects in two clusters.

cluster  $b$  in clustering  $\mathbb{C}$ : about 0.53. To calculate  $P(\omega_1 \cap c_b)$ , each object’s membership of cluster 1 is first multiplied with its membership of cluster  $b$ , resulting in 0.06, 0.16 and 0.63. The mean of these results, about 0.28, is then taken to be  $P(\omega_1 \cap c_b)$ .

$I$  is at its maximum value when clustering  $\Omega$  perfectly recreates  $\mathbb{C}$ . The value stays the same, however, when the clusters in that clustering are split into yet smaller clusters. In other words: positing a new cluster for every node – a meaningless result – will also result in a perfect score. To overcome this problem,  $I(\Omega; \mathbb{C})$  is divided by the average entropy of each clustering. Entropy tends to increase when there are more clusters, resulting in the penalising of undesired behaviour in NMI.

### 2.3.3 Jaccard Similarity Coefficient

To answer sub-question two (“Can topological communities be detected by interpreting textual content?”), Girvan and Newman’s topological community discovery algorithm described in Section 2.2 is used. This algorithm results in a hard clustering (i.e. each object is a member of exactly one cluster). Its results will be compared to those of the application of LDA, which is converted to a hard clustering as well by simply assigning each user to the cluster that user is most affiliated with. When comparing hard clusterings, several other measures are available. For example, a popular one is the Rand Index, which can be defined as in equation 2.7 (Hubert and Arabie, 1985).

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \quad (2.7)$$

Definitions of the symbols used are listed in Table 2.3. The Rand Index counts the number of pairs of objects that are classified the same (i.e. both clusterings consider them to be in the same cluster, or both clusterings consider them to be in different clusters) in both clusterings as a fraction of the total number of pairs of objects.

The problem with the Rand Index, for the intents and purposes of this thesis, is that having a lot of clusters will skew the results. With each object being a member of exactly one of many clusters, the odds of any two objects being

Symbol	Meaning	Definition
$TP$	True positives	The number of pairs of objects of which both objects are placed in the same cluster in both clusterings.
$TN$	True negatives	The number of pairs of objects of which both objects are placed in different clusters in both clusterings.
$FP$ and $FN$	False positives and false negatives	The number of pairs of objects of which both objects are placed in the same cluster in one clustering, but in different clusters in the other.

Table 2.3: Symbol definitions for the Rand Index (equation 2.7).

assigned to a different cluster are very high in both clusterings. This means that the Rand Index will almost certainly end up being close to one.

A more appropriate index is the Jaccard Similarity Coefficient, also called the Jaccard Index, commonly attributed to Jaccard (1901). Unlike the Rand Index, it only considers the overlap in classifications as a fraction of the total classifications, disregarding instances where both clusterings assign a pair of objects to different clusters. Using the same symbols as in equation 2.7, the Jaccard Index is defined in equation 2.8 (Yin and Yasuda, 2006).

$$JI = \frac{TP}{TP + FP + FN} \quad (2.8)$$

As an example, compare the clustering in Figure 2.13 to that in Figure 2.15. There are four pairs of nodes of which both nodes are in the same cluster in both clusterings (so  $TP = 4$ ). However, the one node that switched clusters is part of five pairs (one for each of the other nodes): for all of these pairs, both nodes are in the same cluster in one of the clusterings, but in different clusters in the other (so  $FP + FN = 5$ ). Hence, the Jaccard Index is  $JI = 4/(4+5) = \frac{4}{9}$ .

By not incorporating true negatives, the Jaccard Index evades the problem of the inclusion of many clusters skewing the results. Furthermore, it remains easy to interpret: the resulting number is the fraction of object-pairs that are in the same cluster in one clustering that are also in the same cluster in the other clustering. When both clusterings perfectly predict the other, the Jaccard Similarity Coefficient will be 1.

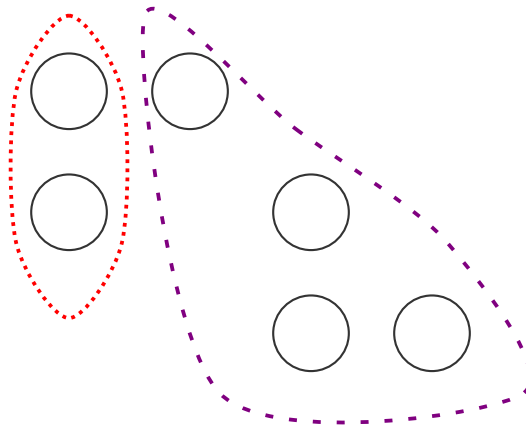


Figure 2.15: Alternative clustering to that in Figure 2.13.

## Chapter 3

# Detecting Communities of Interest

Chapter 2 defined the tools used in this thesis. This chapter and the next use the community discovery algorithms defined there to discover communities on two different data sets, and use the evaluation methods discussed in the previous chapter to compare them to each other and to a gold standard.

This chapter deals with the first sub-question of this thesis:

Can communities of interest be detected by interpreting textual content?

This question is answered as follows. First, a data set is gathered that includes textual content, data about the authors of that content, and actual data on communities of interest that signifies which communities of interest people are a member of, and to what extent – this is the gold standard. Then, using the topic modeling algorithm LDA (see Section 2.1), the textual content is used to train a model to find out on its own what it thinks the composition of the communities of interest is. Finally, the predicted composition is evaluated against the actual composition of the gold standard using Normalised Mutual Information (see Section 2.3.2).

This chapter starts with a discussion of the source of the gold standard and the text to train on. This is followed by a number of statistics that provide a high-level overview of the size and composition of the gold standard. After that there is a description of the discovery of communities by LDA based on the textual content of the data source. The chapter ends with an evaluation of the results of the experiments: how well were the actual communities of interest predicted?

### 3.1 reddit

The source of the data used in this chapter is the social news and entertainment website **reddit**<sup>1</sup>. In a nutshell, this website enables registered users to post links to interesting stories or to write their own content. These stories can

---

<sup>1</sup><http://reddit.com/>

receive positive and negative votes (upvotes and downvotes, respectively, in reddit lingo), that respectively positively and negatively influence the post’s likelihood of being shown to visitors. This results in a collection of links with, at least theoretically, the best links displayed first. Hence, reddit calls itself “the front page of the internet”.

Content posted to reddit is organised in so-called *subreddits*. A subreddit is nothing more than a label to categorise content. Provided that a certain label, stylised as `/r/<labelname>`, still is available, any user can create a subreddit for other users to submit content to. Whenever a user submits content (i.e. a link, or textual content – a *self post* in reddit lingo), it is submitted to one of the seemingly infinite subreddits (one such subreddit, `/r/ofcoursehatsathing`, is dedicated to linking to other subreddits that you would not expect to exist). This results in subreddits being a collection of content related to the same topic. An example is shown in Figure 3.1, which shows the first page of `/r/sixwordstories`, where users share stories consisting of a mere six words.

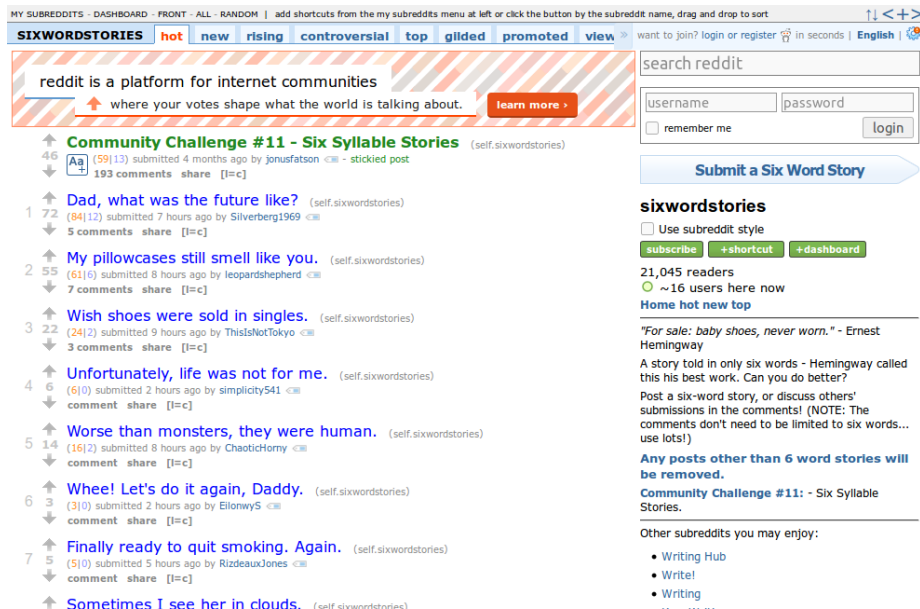


Figure 3.1: Example subreddit: `/r/sixwordstories`

Key to reddit are the upvotes and downvotes. Each user can either upvote or downvote a post, resulting in a score that show the website visitors’ appreciation of the content. Figure 3.2 shows the upvote and downvote arrow surrounding the content’s score, with the upvote button having been pressed, while Figure 3.3 depicts a similar situation but with the downvote button having been pressed.



Figure 3.2: Example of an upvote





Figure 3.3: Example of a downvote

Individual subreddits can be sorted on the highest rated posts for a given period of time, but users can also visit `/r/all`, which aggregates content from all subreddits based on popularity. Figure 3.4 shows the top posts of all time, up to the time of writing. Posts here come from a variety of subreddits, such as `r/pics` and `/r/IAmA`. Likewise, but subtly different, the front page of reddit consists of an aggregate of posts from subreddits the user has subscribed to, and will thus show the most popular posts related to subjects of the user's interests.

The screenshot shows the Reddit front page for `/r/all`. The top navigation bar includes links for 'ALL', 'hot', 'new', 'rising', 'controversial', 'top', 'gilded', and 'view images (17)'. Below the navigation bar, there is a search bar and a 'login' button. The main content area displays a list of posts, each with a score, a title, a submission time, and a comment count. The top post is a sponsored link for a shirt giveaway. Other notable posts include 'test post please ignore', 'The Bus Knight', 'I live in the same valley as Adam West...', 'TIL that hitting the upvote arrow gives you reddit gold for free', 'This guy is a reporter on Fox 2 here in Detroit...', 'After searching FB for people with the same name as me...', and 'I am Barack Obama, President of the United States -- AMA'. On the right side, there is a sidebar with options like 'Use subreddit style', 'Exclude your subscribed subreddits', and 'See gilded comments and submissions'. At the bottom right, there is a 'daily reddit gold goal' progress bar showing 69% completion.

Figure 3.4: Top posts on `/r/all` of all time

Another characteristic part of reddit is the vibrant comment section. Users can comment on each post, and reply to other comments. Just as with links, users can upvote and downvote comments, leading to the most appreciated comments to surface to the top. An example of a comment and replies are in Figure 3.5.

reddit comments are a valuable data source for this research. They have several advantages. First of all, there are a lot. On their blog, reddit's owners estimate that its users left about 260 million comments in 2012 alone<sup>2</sup>. This means that it is relatively easy to compile a dataset large enough to meaningfully

<sup>2</sup><http://blog.reddit.com/2012/12/top-posts-of-year-and-best-of-2012.html>.



Figure 3.5: A series of comments

apply LDA to.

Secondly, reddit's comments are easily accessible programmatically. It is supported and even encouraged to create third-party applications that interface with reddit, under certain conditions (the most important of those being to not overwhelm the site with content requests). Accessing content that requires user authorisation is documented at <http://www.reddit.com/dev/api>. Additionally, content that is publicly visible can be accessed in machine-parseable formats by appending `.json` or `.xml` for the JSON or XML formats, respectively, to that page's URL.

Finally, since comments are related to posts, and posts are grouped according to topic (subreddit), it makes for an ideal gold standard. Commenting on posts involving a certain subject can be seen as taking part in a community finding common ground in that subject, hence representing communities of interest.

Of course, there is no such thing as a free lunch. Not all information about reddit users is publicly disclosed; most crucially, the communities they are subscribed to is beyond limits. Although it is not required to be subscribed to a subreddit to comment there, the fact that a user made the effort to comment on a post in a certain subreddit implies that this user feels part of it.

Furthermore, comments on reddit can be as long or as short as you like. And of course, it is completely optional to post comments. This means that not all users might have produced sufficient text to classify them.

## 3.2 The gold standard

In order to be able to assess whether we can detect communities of interest on the basis of language, we need a gold standard. It can provide information on what communities of interest are actually present. It should be the definitive authority on which users belong in a community together. In other words: when LDA determines which users are likely to be interested in the same subject, there should be a way to determine whether that is actually the case. Thus, the data set should exhibit the following properties:

1. It should encompass a large number of users, including a large amount of text posted by each user.
2. For each user, it should be known what communities that user participates in most.

The first step in gathering the data is determining which users to investigate. To make sure each user still identifies with the same communities, only users who have recently been active are considered. To ensure that the data set does not only include users from specific time zones, a list of all users that have left at least one comment in a 24 hour period is compiled as a starting point of the data gathering.

Then, the last 100 comments by these users are fetched, together with the associated metadata. Most important of this metadata, apart from the name of the author of the comment, is which subreddit the comment has been placed in. For each of these subreddits, the percentage of the user's comments that are posted to it are taken to be that user's ratio of participation in and commitment to that community. This will form the gold standard of the user's community membership.

Finally, this one large dataset is split in ten smaller datasets. Users were sorted alphabetically, and each set  $x$  consists of the comments by every  $x^{th}$  user in the list. This results in ten different datasets with similar characteristics, allowing the experiment to be repeated ten times, with ten different gold standards to compare to.

To give an idea of the composition of the data, Table 3.1 lists the number of comments included in each of these datasets, and the number of different subreddits these comments were posted in. Various statistics on each dataset are listed in Tables 3.2 to 3.6. As can be seen, the dataset is vast, spanning a large number of users writing a large number of words in a large number of subreddits. Table 3.2 shows that subreddits can be home to a lot of comments, but there is also a large number of subreddits that is home to just few comments - this is in line with the low number of different users that many subreddits have, as shown in Table 3.5. Table 3.3 shows that most users post a lot of comments. Most of these comments are more than one-liners, as Table 3.6 shows: most comments contain tens of words. Table 3.4 shows that users largely confine themselves to about 12 subreddits to participate in.

Dataset	# comments	# subreddits
1	74009	2774
2	72629	2674
3	75375	2702
4	74958	2715
5	71174	2575
6	75775	2686
7	74960	2658
8	71805	2609
9	74514	2737
10	74111	2672

Table 3.1: Number of comments included in each dataset

Dataset	Minimum	Median	Mean	Maxium
1	1	3	26.68	9572
2	1	3	27.16	9045
3	1	3	27.9	9921
4	1	3	27.61	9180
5	1	3	27.64	8744
6	1	3	28.21	9972
7	1	3	28.2	9441
8	1	3	27.52	9305
9	1	3	27.22	9737
10	1	3	27.74	9148

Table 3.2: Number of comments per subreddit in each dataset

Dataset	Minimum	Median	Mean	Maxium
1	1	55	56.54	100
2	1	53	55.48	100
3	1	57	57.58	100
4	1	59	57.26	100
5	1	49	54.33	100
6	1	59	57.84	100
7	1	56.5	57.22	100
8	1	51	54.81	100
9	1	55	56.92	100
10	1	56	56.62	100

Table 3.3: Number of comments per user in each dataset

Dataset	Minimum	Median	Mean	Maxium
1	1	11	12.75	78
2	1	11	12.67	72
3	1	12	13.01	50
4	1	12	12.81	83
5	1	11	12.1	44
6	1	12	12.97	49
7	1	11.5	12.85	48
8	1	11	12.58	71
9	1	11	12.87	67
10	1	12	12.52	81

Table 3.4: Number of subreddits users participate in in each dataset

Dataset	Minimum	Median	Mean	Maxium
1	1	1	6.018	849
2	1	1	6.202	823
3	1	1	6.301	858
4	1	1	6.178	866
5	1	1	6.157	823
6	1	1	6.324	872
7	1	1	6.332	874
8	1	1	6.318	872
9	1	1	6.157	847
10	1	1	6.132	860

Table 3.5: Number of users participating in subreddits in each dataset

Dataset	Minimum	Median	Mean	Maxium
1	6	39	73.07	3383
2	4	38	69.75	3282
3	6	38	69.27	4006
4	4	38	69.1	3262
5	4	38	71.06	3784
6	4	38	72.17	3508
7	4	38	72.7	3075
8	4	38	70.92	4421
9	4	38	71.15	6272
10	4	40	73.34	3545

Table 3.6: Number of words per comment in each dataset

### 3.3 Detecting communities

As described in Section 3.2, for all users that have posted in a 24 hour period, the 100 latest comments they have placed have been gathered. Since the goal is to detect communities of interest based on user text alone, all metadata except for a comment's author is ignored for the community detection. The contents of each comment, however, is processed so as to remove all formatting and links, the result of which is concatenated to all other comments by the same user into a single file. This latter step is crucial in making LDA-discovered topics categorise users into communities of interest; otherwise, it would be classifying individual documents. An additional advantage, as found by Hong and Davison (2010), is that the aggregation of all comments by the same user improves the quality of the model.

For each dataset, LDA was run with the users' concatenated posts as input as described in Section 2.1. The number of communities was set to be equal to the total number of subreddits that were represented in the gold standard – automatically determining the optimal number of communities is left as an open research question, and is further elaborated upon in Chapter 5. The implementation used is the one from the MALLET toolkit (McCallum, 2002). MALLET is an open source software package including many tools for statistical language processing, including an implementation of LDA and tools for e.g. the removal of stop words. MALLET's default hyperparameter values were used due to the heuristical nature of these parameters: a cumulative value of 50.0 for the  $\alpha$  parameter vector<sup>3</sup> and 0.01 for all values of the  $\beta$  parameter vector. As described in Section 2.1, these values define the concentration parameters of the Dirichlet distributions. Since these values are small, documents are likely to be about few topics, and topics are likely to be characterised mostly by only a small percentage of the words in the vocabulary.

Table 3.7 shows a subset of the community memberships of one dataset detected by LDA to give some insight in the composition of the data. For example, the probability of a comment by user `Mrwhitepantz` to be in community 203 is, as expected by the output of this run of LDA, 0.112.

Table 3.8 shows the top words for five randomly selected communities as detected in one of the experiments. Intuitively, community #16 could be argued to be about gaming (like `/r/gaming`), community #70 might be about Canada (like `/r/canada`), community #83 could be interested in laws (like `/r/law`) or perhaps in gender equality (like `/r/feminism` or `/r/MensRights`). Community #112 might have a common interest in the internet or internet-related news (like `/r/mozilla`), and community #124 might be very interested in social networks/reddit itself (like `/r/circlejerk`).

Qualitatively, the topics of communities some users were assigned to appeared to be relatively accurate, after superficially scanning the comments they left and the subreddits they were active in. For example, user `FasterThanTW` was most strongly connected to community #16, and has commented in subreddits such as `/r/gaming`, `/r/wiiu` and `/r/Games`, among others. On the other hand, though, user `CornealRefraction` is strongly linked to topic #70 (which seems related to Canada), but almost all of this user's comments are in

---

<sup>3</sup>Since this value is cumulative, the values of the individual parameters in the vector are 50 divided by the number of topics.

User	Comm. #1	Comm. #1 par- ticipation probabil- ity	Comm. #2	Comm. #2 par- ticipation probabil- ity	...
Mrwhitepantz	203	0.112	186	0.101	...
hurf_mcdurf	159	0.088	104	0.084	...
gladdo420	101	0.128	153	0.111	...
annasbadman	208	0.891	159	0.015	...
⋮	⋮	⋮	⋮	⋮	⋮

Table 3.7: Some of the community membership percentages for some of the users in one of the datasets.

16	70	83	112	124
games	buffalo	wrong	service	reddit
ps	city	isn	mozilla	op
game	yeah	argument	battery	source
pc	pretty	law	product	comment
console	cities	aren	charging	post
buy	eggs	thing	firefox	photo
xbox	montreal	rights	isp	picture
nintendo	cool	child	mana	joke
pokemon	blood	claim	level	facebook
gen	quebec	laws	hero	youtube
titanfall	byzantine	population	trademark	downvoted
fps	nyc	children	lion	nope
wii	cat	agree	distribution	posting
consoles	toronto	conversation	mah	fake
gaming	york	majority	transit	dog
developers	california	statement	installation	dude
controller	park	men	dell	attention
save	canadian	shouldn	lane	website
bought	club	rape	marks	karma

Table 3.8: Top words for five randomly selected communities

/r/WTF, a subreddit for sharing bizarre phenomena and events or anything else out-of-the-ordinary.

However, we would like to quantify how much the detected communities align with the actual subreddit composition. Since users can participate in multiple subreddits, and LDA assigns users to different communities proportionally, these clusterings are said to be *soft clusterings* (see Section 2.3). Furthermore, since subreddits do not contain other subreddits, and LDA's communities are non-hierarchical as well, the clusterings are also *flat* (again, see Section 2.3). The applicable method, in this case, is Normalised Mutual Information. For each of the experiments, NMI was used to perform two comparisons:

- LDA-based community assignments to actual subreddit participation
- Randomly assigned community participation to actual subreddit participation

The first comparison quantifies the amount of information the LDA-based community assignments provide about users' subreddit participation. The closer this number is to one, the better the latter can be predicted by the former, or in other words: the better both assignments match. The second comparison quantifies the amount of information about the actual community participation is provided by randomly assigning users to communities (this should be close to 0).

Unfortunately, NMI by itself cannot be used to make definitive statements regarding e.g. the number of users clustered correctly. However, it is a useful tool for benchmarking the performance of different algorithms. Since there do not appear to be other algorithms for finding communities of interest to compare to, performance will be assessed by comparing it to an algorithm that randomly assigns users to specific communities of interest. Thus, the output will be similar to that in Table 3.7, but with the community numbers and their respective participation probabilities chosen at random. The expectation then is that LDA will consistently perform better.

The results of the comparisons for each of the ten datasets can be seen in Table 3.9. As can be seen, it is the case that predicting the community assignments using LDA consistently outperforms randomly assigning users to communities.

There was a desire to further rule out that the clustering prediction by LDA outperforming the random clustering prediction was by chance. To do this, for one of the experiments, 204 additional random clustering predictions were done. Their NMI, indicating how well they predicted the gold standard, was plotted. If LDA's prediction of the gold standard really is better than a random prediction, it should fall in the 95th percentile of the density function (which shows which NMI's occur most often). The density function of the NMI of the random predictions is plotted in Figure 3.6, with the 95th percentile shaded. Since the LDA-based prediction resulted in a rounded NMI of 0.1003, this is clearly the case: it was two orders of magnitude better than 100% of the random predictions.



LDA-actual	Random-actual
0.4793	0.0040
0.3183	0.0051
0.1995	0.0046
0.3667	0.0042
0.2583	0.0038
0.3514	0.0050
0.2054	0.0038
0.1821	0.0040
0.2488	0.0040
0.1003	0.0013

Table 3.9: Rounded NMI for ten groups of user comments.

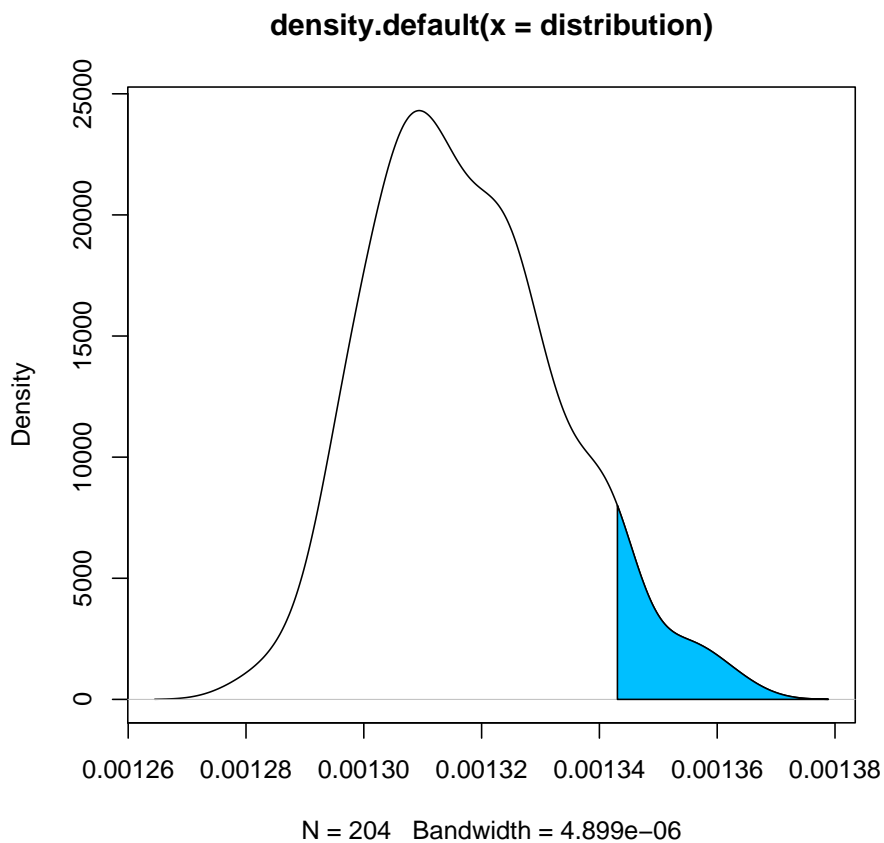


Figure 3.6: Density function of the NMI of 204 random predictions of the community composition of the gold standard. The 95th percentile is shaded.

### 3.4 Conclusions

LDA is an interesting tool to detect communities of interest. It requires no specific user action other than simple participation. Not only does it provide information on which users share interests, but it also offers insight into what these interests entail.

Of course, these features are only useful if this information is meaningful. It is difficult to make definite statements, considering the unavailability of performance measures that can tell us e.g. how far off exactly a clustering method is from the gold standard. What is available, though, is a method of comparing different clustering algorithms. Using this method, it can be seen that LDA provides substantially more information about the composition of the gold standard clustering than a random method. Thus, we can conclude that it is at least somewhat meaningful. Furthermore, it is now available as a benchmark for other methods of predicting communities of interest to compare to.

Returning to the first sub-question:

Can communities of interest be detected by interpreting textual content?

It can be said that it is possible to some extent, although it is hard to determine to which extent exactly. In any case, a benchmark is now available, allowing for the possibility of seeing whether performance can be improved.

## Chapter 4

# Comparison to topological communities

The previous chapter attempted to provide an answer to the first sub-question of this research. To answer the main research question, the second sub-question will have to be answered first – that will be the focus of this chapter.

The second sub-question of this thesis was:

Can topological communities be detected by interpreting textual content?

Unfortunately, to answer this question, the reddit dataset from Section 3.2 cannot be re-used: there are no publicly accessible user-user-relationships on reddit that can be used to construct communities based on user topology. Therefore, to answer this question, another dataset will be used that contains both person-person-relationships and human-generated textual content. On this dataset, communities of interest will be detected using LDA, in the same way as was done in Section 3.3. Additionally, the Girvan-Newman algorithm for topological community discovery from Section 2.2 will be applied to the relationships between people in this dataset. The Jaccard Index, as described in Section 2.3.3, can be used to calculate the extent to which these different types of community agree on the clustering of the people in the dataset.

This chapter is structured as follows. The first section contains a description of the Enron email corpus, the data source that contains both person-person-relationships and human-generated textual content. This is followed by a section containing the details of the topological community detection on the included network, and an overview of the structure of those communities. The same is then provided for the detected communities of interest. Finally, there is an evaluation of how well these two clusterings align, and conclusions that can be drawn from that.

### 4.1 Enron

Enron Corporation was an energy company that went bankrupt in 2001 after it was uncovered that the company was guilty of large-scale accounting fraud.

One of the side effects of the Enron scandal, as it came to be known, was that a large database of employees' emails were eventually acquired and made public, providing the research community with a large corpus of real emails by real people, unbound by legal restrictions. This results in a corpus that satisfies the following useful properties:

1. It includes a large number of people that have each produced large bodies of text on which LDA can be applied to discover communities.
2. It entails a large number of people that can be connected in a network: two persons can be said to be related when communication has taken place between them. This can be used to discover topological communities.

The original corpus includes 619446 messages by 158 Enron employees (Klimt and Yang, 2004). However, the data is preprocessed to prepare it for the application of LDA to detect communities of interest. First of all, it is cleaned up to remove email headers and other noise such as signatures<sup>1</sup>. Then, for all employees, all emails sent by that employee are concatenated into a single document, as described in Section 3.3. Furthermore, since its first publication, several employees' emails have been removed from the dataset by the maintainers as per those employees' requests. All in all, in total the final dataset of emails consists of the concatenated emails of 148 employees, with an average size of about 550kB.

## 4.2 The topological network

The topological network is constructed by assuming that two Enron employees are related when an email was sent between them. It is created using a script written in the GNU R programming language that

1. Loads a prepared import<sup>2</sup>.
2. Finds all emails sent by an Enron employee to another Enron employee.
3. Parses this into an adjacency matrix, drawing an edge between two employees whenever at least one email had been sent by one to the other.
4. Feeds the adjacency matrix to the igraph library<sup>3</sup>.
5. Uses igraph's implementation of the Girvan-Newman algorithm to create multiple clusterings with an increasing number of clusters.
6. Returns the clustering that igraph determined to have the highest modularity (see Section 2.2).

Figure 4.1 shows the employee graph with the employee nodes coloured according to the communities detected using the above method. Each employee can only be a member of one community. Modularity was highest when the employees were clustered in 83 communities, so that is the number of clusters included.

<sup>1</sup>Based on a script published at <http://java.dzone.com/articles/topic-modeling-python-and-r-0> (retrieved on the 29th of June, 2014).

<sup>2</sup>The import was obtained from <http://www.ahschulz.de/enron-email-data/>.

<sup>3</sup>For more information, see <http://igraph.org/>.

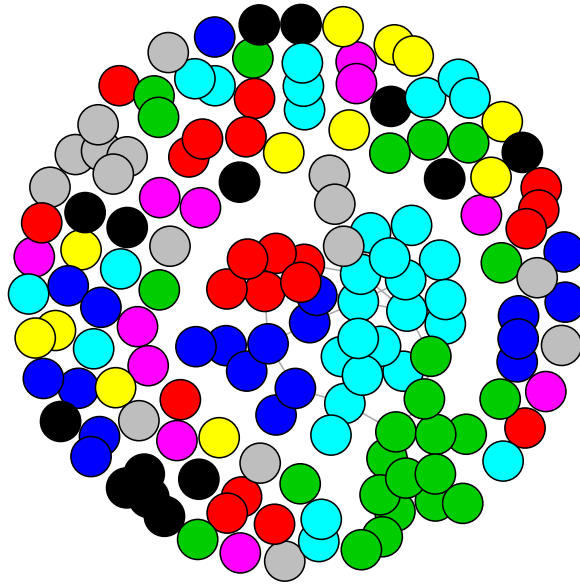


Figure 4.1: Graph of Enron employees – adjacent nodes with the same colour are from the same detected community. Note that some colours are repeated, as there are more communities than colours (namely 83).

### 4.3 Communities of interest

To enable comparison to the output of the Girvan-Newman algorithm, LDA is set to find 83 communities as well – if the types of discovered community indeed have a similar structure, then the number of communities should match as well. Just as in Section 3.3, the MALLET software package was used with its default values for the hyperparameters. Additionally, only the community an employee is most strongly assigned to is considered. This is to ensure that each employee is a member of a single community only, just as was the case when the communities were detected using the Girvan-Newman algorithm.

Table 4.1 shows the top words for five randomly selected detected communities. The communities seem to have relatively distinct subjects of conversation, especially considering how the number of communities was decided entirely by the number of communities detected using the Girvan-Newman algorithm for the run with the highest modularity. There are also a few words like “forwarded” and “filename” which might have been used by employees, but might also have been missed by the cleanup script. Unfortunately, it is practically impossible to completely and correctly clean a large corpus like this.

If the communities detected by LDA are similar to those detected by the Girvan-Newman algorithm, that would be a win: apart from finding the communities, there would also be information about the particular interests of those communities. If they differ greatly, though, it will be instructive for future research to keep in mind that topological communities might not always directly represent communities of interest, and drawing conclusions from them other than about the ties between people might not be justified.

As discussed in Section 2.3.3, the Jaccard Similarity Coefficient is a good way of measuring the similarity of to flat, hard clusterings like these (Pang-Ning et al., 2006). The Jaccard Similarity Coefficient of the Girvan-Newman based clustering and the LDA-based clustering was **0.04477612**. This means that of all pairs of employees that are placed in the same community in either of the clusterings, only about 4.5% were placed in the same community in *both* clusterings.

A closer inspection of Figure 4.1 indicates that the clustering with the highest modularity still contains a lot of communities containing just a single employee. Likewise, 83 different communities of interest intuitively feels like a lot for a single company. Therefore, to find out whether this was the reason for the low the similarity between the two clustering methods, the experiment was repeated with the following changes:

- Employees who were the only member of their communities in the output of the Girvan-Newman algorithm were removed from the dataset. This resulted in only 20 communities remaining, shown in Figure 4.2.
- On the emails of the remaining employees, LDA was once again applied, but this time set to find only 20 communities. The top words for five of these are listed in Table 4.2.

The communities detected through LDA still seem relatively focused, topic-wise, and the topological communities also seem more convincing. What counts, however, is the new value of the Jaccard Similarity Coefficient. This is, for the pruned employee list, 0.08283133. Although a significant improvement, the

19	24	44	50	74
dana	mmbtu	robin	zufferli	enron
davis	enron	rodrigue	john	time
enron	deliveries	management	filename	pmt
subject	capacity	ectcc	mail	back
forwarded	tw	ees	ca	message
pm	pg	subject	original	day
na	gas	book	johnsubject	ll
denise	california	na	july	give
comcc	lokay	id	canada	today
email	san	positions	jpg	call
aol	michelle	gas	alberta	great
amto	juan	agg	power	morning
yahoo	average	ll	ab	ve
god	large	portfolio	chris	hope
moore	pkgs	dt	pst	week
corp	transwestern	gabriel	rob	long
october	averaged	monroy	amto	send
lisa	sj	position	doc	things
home	ca	communica- tions	calgary	talk

Table 4.1: Top words for five randomly selected communities

value is still low, meaning it is unlikely that this was the root cause for the minimal overlap.

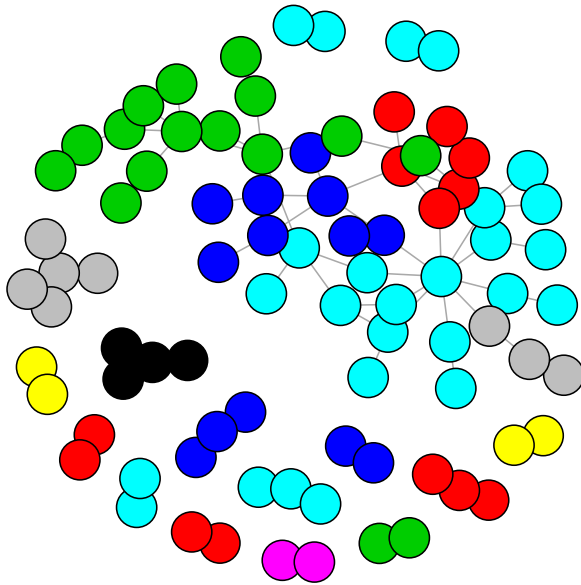


Figure 4.2: Graph of Enron employees and their Girvan-Newman-detected communities, with singleton communities and their sole members removed.



5	7	10	13	18
california	sara	enron	night	nnbsp
jeff	enron	ena	good	gt
power	corp	pm	don	gas
state	agreement	delainey	mail	br
energy	houston	business	weekend	mike
pm	eb	deal	time	west
market	isda	power	ll	socal
mail	america	dave	susan	index
utilities	shackleton	meeting	back	day
customers	fax	team	message	project
davis	north	market	email	file
electricity	smith	john	msn	attached
commission	street	gas	work	capacity
edison	phone	management	day	desk
governor	credit	trading	hope	williams
contracts	doc	year	things	information
puc	texas	project	tonight	paso
utility	attached	mark	love	el
iso	pm	mm	call	pg

Table 4.2: Top words for five randomly selected communities of the twenty communities that remained after purging singleton communities.

## 4.4 Conclusions

The Enron email corpus is a useful dataset in that it provides both relationships between different people, as well as textual content produced by those people. A traditional, topological community detection method found 83 communities of differing sizes, based on who emailed whom. The method proposed in this thesis – repurposing LDA for detection of communities of interest – was configured to find an equal number of communities. If the communities found by these methods were similar in structure (presumably, they would be formed around common interests), the Jaccard similarity coefficient would return a value close to 1. In practice, however, it turned out that only about 4.5% of pairs of employees were clustered together by both algorithms. This casts the shadow of doubt over a commonly made assumption that topological communities represent communities of interest, and suggests that ties between people are not necessarily the result of a common interest.

It is instructive to take another look at the second sub-question:

Can topological communities be detected by interpreting textual content?

The above results give reason to think that, no, community detection through topic modeling is not necessarily a good way of finding the same communities as topological methods do. Assuming that communities of interest are correctly detected by using LDA (see Chapter 3), the communities detected by topological methods are of a different type of community than communities of interest.



## Chapter 5

# Conclusions and Future Work

Social networks are graphs of nodes representing people, the relationships between whom are indicated by drawing edges connecting their respective nodes. These graphs of people are often used to find communities based on their topology by means of discerning clusters of nodes that are densely connected to one another, but are only sparsely connected to nodes outside of their clusters. While community structure is a useful concept for graph theory, it is unclear how it should be interpreted. This thesis attempted to shed some light on the meaning of communities as used in graph theory by comparing them to the concept of communities of interest. It did so in two steps.

First, in Chapter 3, a method was proposed to automatically detect communities of interest based on textual content. The method entails repurposing the topic modeling algorithm LDA for the discovery, not of the topics of documents, but of the communities of interest that people belonged to. The algorithm was then tested against a newly constructed gold standard based on social news website *reddit*, which contains user-defined groups that are taken to be the definite authority on the composition of communities of interest. Although concrete conclusions are difficult to draw, the results showed that communities discovered through the use of LDA gave a significantly better indicator of the composition of the communities of interest than random guesswork.

Chapter 4 then compared the communities discovered using methods based on a graph's topology to communities of interest in order to find out whether the two types of community clusterings overlapped. Since no easily accessible dataset was found that included both person-person-relationships (for constructing a graph) *and* information on the actual composition of communities of interest, the LDA-based community detection method was used to find communities as a proxy for communities of interest. Both this method and Girvan & Newman's algorithm for topologically detecting communities were applied to the Enron email corpus. The resulting two clusterings differed greatly, hinting that the communities often discussed in the literature dealing with social networks can not simply be assumed to be communities united by joined interests.

These two steps answered the sub-questions presented at the start of this thesis. The answers then lead to the answer of the main research question:

Do communities of interest align with topological communities?

Since topic modeling appears to give a reasonable indication of the composition of communities of interest when compared to a random estimation of their composition. As as of yet there are no other methods of community detection that it can be compared to, it is currently the most reliable method of finding communities of interest. When taking it as a reasonable indication of the composition of communities of interest, the fact that the communities detected by it seem to differ greatly from the topologically detected communities leads to the conclusion that communities of interest do not align with topological communities.

In answering this question, this thesis provided the following main contributions. First of all, it shed more light on the structure of topological communities, testing whether they can be considered to be what is understood as communities of interest. It also proposed a method of discovering those communities of interest, describing how to repurpose an existing topic modeling algorithm for this new use case. The thesis also provided a clear methodology of how to evaluate algorithms for discovering communities of interest, and a benchmark to compare the performance of these algorithms. Along with this benchmark, a data source has been found (reddit) that provides both a gold standard and data to train potential algorithms on. This thesis contains instructions on how to collect the data, and Section 5.1 outlines the steps that should be taken to create a more definitive dataset that could be used in future research. Finally, this thesis hopefully brought into the limelight some implicit assumptions in the discussion of community structures in social graphs, and created awareness about what those community structures could actually represent in human relationships.

The rest of this chapter will deal with potential improvements to this research that were not implemented due to lack of time, and will suggest new avenues for future research with high potential.

## 5.1 Improve the gold standard

While sufficient for this research, the reddit-based gold standard leaves room for improvement that can be utilized to improve it for future research.

First of all, although a lot of work would have to be carried out, manual curation of a list of subreddits that are clearly focused on single topics could result in a more convincing gold standard. For example, the data set used in this thesis included subreddits such as `/r/pics`, `/r/gifs` and `/r/videos` that contain wildly different types of content with the only common denominator being the medium they are provided in, and `/r/news` focuses on news about any given topic. A disadvantage is that more specific subreddits are often less popular, meaning that data will likely have to be gathered about a larger period of time, increasing the probability that users only participated in that community for parts of that period.

Another, less labour-intensive, method of improvement would be to instate a member threshold for communities. Table 3.5 shows that many of the subreddits only exist to facilitate a single user (likely the creator), and that most subreddits only include a few comments to begin with (Table 3.2). It is not hard to argue

that a single user does not constitute a community. Therefore, one could be justified to only consider communities that have at least a certain number of users actively participating in it, discarding the long tail of small communities.

Finally, a very obvious improvement, but one that is only available to the reddit administrators, would be to incorporate the actual subscription data by users. That is, users can decide of which subreddits they want the posts to be on their front page when they are logged in to reddit. Presumably, this would be an even better indicator as to with which subreddits users identify – after all, they have indicated so themselves. A disadvantage would be that all subscriptions are equal, in other words: they provide no information about which subreddits users deem most relevant to their interests.

Apart from improving the quality of the gold standard, it would also be helpful to have a benchmark of how well humans can predict users' membership of communities of interest on reddit. If the number of different communities would be limited, a service like Mechanical Turk<sup>1</sup> could be used to classify a large number of users. This would allow researchers to find out whether their algorithms can match human performance, instead of only comparing them to other algorithms.

## 5.2 Improve the quality of the data sets

The bodies of text used to discover communities of interest among people were of sufficient quality for this thesis, but they could still be better. For example, while most of the users included in the reddit dataset posted a lot of comments, accumulating large bodies of text, this was not the case for all users. To illustrate this point, consider Table 3.3, repeated here as Table 5.1. It shows that there are users that have only posted a single comment. It is unlikely that a single comment provides meaningful information about the full span of that user's interests, making it difficult for a human, let alone an algorithm, to find out which communities such a user should belong to.

The textual content itself, although largely cleaned up, still showed some signs that some noise remained. The words often used in certain communities

<sup>1</sup>See <https://www.mturk.com/>

Dataset	Minimum	Median	Mean	Maxium
1	1	55	56.54	100
2	1	53	55.48	100
3	1	57	57.58	100
4	1	59	57.26	100
5	1	49	54.33	100
6	1	59	57.84	100
7	1	56.5	57.22	100
8	1	51	54.81	100
9	1	55	56.92	100
10	1	56	56.62	100

Table 5.1: Number of comments per user in each reddit dataset

detected on the Enron corpus, for example, included some words often used to describe an email (see Table 4.1). This indicates that the cleanup script might not have fully dealt with all the different email clients' ways of including metadata in emails. On reddit, users can quote one another or sources they mention. Although this presumably only comprises a small percentage of a user's accumulated comments, for optimal representation of the user's interests it would be best to remove these comments.

Alternatively, one could try to find entirely different sources for the data sets. Apart from the data sets being difficult to clean up, one could argue that email contacts are simply not a good indicator of a relationship between people. Ideally, an alternative data set includes real person-person relationships, human generated textual content, and a gold standard of the composition of communities of interest. While these seem like high demands, Facebook<sup>2</sup> is an example of a website where

1. It is common for users to “friend” each other, often reflecting real-life friendships. This allows one to construct a social graph of the users.
2. Users produce large amounts of texts over the course of their membership, be it in comments or in so-called “wall posts” on their own profile.
3. Many users are part of communities of interest through group memberships and “likes”.

Unfortunately all this data was not available from Facebook and similar websites for this thesis. For whom this data is within reach, however, it would be an interesting fit for this research.

### 5.3 Incorporate LDA research

LDA is a popular algorithm, meaning that a sizable amount of research has been performed using and extending it. It also led researchers to try to find ways to improve the performance of the algorithm. The use of LDA as the algorithm for finding communities enables the potential use of that research to improve upon the results of this thesis.

Some research focuses on relaxing LDA's assumptions. Many of these relaxations might apply not just to document clustering, but also to the discovery of communities of interest. For example, the Correlated Topic Model (CTM) by Lafferty and Blei (2006) assumes that topics are related: documents about life sciences are more likely to also be about biology than about Christianity. It is not unlikely that the same holds true for communities of interest: people in a community interested in machine learning are more likely to also be a member of a community about Linux than a community about the Netherlands. Using CTM instead of LDA might just improve the results.

Another area where LDA is believed to leave room for improvement is its parameters. With plain LDA, the person running it has to decide, mostly through guesswork, how many topics (or communities) a corpus is to encompass. In this research, this was not that much of a problem: the number was simply set to be the same as the number of communities in the clusterings it was compared

---

<sup>2</sup>See <https://facebook.com/>

to, be it the gold standard or a topological clustering. Ideally though, a good community detection algorithm should run without the help of a gold standard telling it how many communities there are. An example of an algorithm that attempts to accomplish this is hLDA by Griffiths and Tenenbaum (2004).

## 5.4 Incorporate other features

Although, in this thesis, LDA groups people based on the words they use, the meaning or order of these words hardly matter. Stopwords are removed, but apart from that, all words are equal. For this thesis, this was sufficient. However, one could think of interesting applications that could take advantage of other properties of language. For example, one could try to take into account the number of first-person pronouns to measure a person's commitment to a community. Other examples of potential avenues of research include the use of sentiment analysis to measure the mood of a community, the use of a person's mastery of community specific terms as an indicator for this person's centrality within the community, or taking into account the use of different languages in certain communities.

Metadata that is available could also point to interesting avenues for future research. An obvious example is the time at which certain text was published. This could be used to find out, for example, whether the interests of a community changes over time. This could, in turn, take into account the members of the community, and whether perhaps a change in a person's language use can predict a change in a person's community membership, or that perhaps people in a community influence each other enough to change a community's interest as a whole.





# Bibliography

- Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 44–54, New York, NY, USA, 2006. ACM. ISBN 1-59593-339-5. doi:10.1145/1150402.1150412. URL <http://doi.acm.org/10.1145/1150402.1150412>.
- David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012. ISSN 0001-0782. doi:10.1145/2133806.2133826. URL <http://doi.acm.org/10.1145/2133806.2133826>.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=944919.944937>.
- Danah Boyd. *Faceted id/entity: Managing representation in a digital world*. PhD thesis, Massachusetts Institute of Technology, 2002.
- Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *International AAAI Conference on Weblogs and Social Media*, 2011. URL <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847/3275>.
- Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 307–318, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-2035-1. URL <http://dl.acm.org/citation.cfm?id=2488388.2488416>.
- Ying Ding. Community detection: topological vs. topical. *Journal of Informetrics*, 5(4):498–514, 2011.
- Neveen Ghali, Mrutyunjaya Panda, Aboul Ella Hassanien, Ajith Abraham, and Vaclav Snasel. Social networks analysis: Tools, measures and visualization. In *Computational Social Networks*, pages 3–23. Springer, 2012.
- M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826,

2002. doi:10.1073/pnas.122653799. URL <http://www.pnas.org/content/99/12/7821.abstract>.
- DMBTL Griffiths and MIJJB Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16:17–24, 2004.
- Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.
- Tom Griffiths and Mark Steyvers. *Latent Semantic Analysis: A Road to Meaning*, chapter Probabilistic topic models. 2006.
- France Henri and Béatrice Pudelko. Understanding and analysing activity and learning in virtual communities. *Journal of Computer Assisted Learning*, 19(4):474–487, 2003.
- Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.
- Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. ISSN 0176-4268. doi:10.1007/BF01908075. URL <http://dx.doi.org/10.1007/BF01908075>.
- Paul Jaccard. étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, pages 547–579, 1901.
- Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *Machine learning: ECML 2004*, pages 217–226. Springer, 2004.
- Robert V Kozinets. E-tribalized marketing?: the strategic implications of virtual communities of consumption. *European Management Journal*, 17(3):252 – 264, 1999. ISSN 0263-2373. doi:10.1016/S0263-2373(99)00004-3. URL <http://www.sciencedirect.com/science/article/pii/S0263237399000043>.
- JD Lafferty and MD Blei. Correlated topic models. In *Advances in Neural Information Processing Systems, Proceedings of the 2005 conference*, pages 147–155, 2006.
- Daifeng Li, Bing He, Ying Ding, Jie Tang, Cassidy Sugimoto, Zheng Qin, Erjia Yan, Juanzi Li, and Tianxi Dong. Community-based topic modeling for social tagging. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1565–1568. ACM, 2010.
- Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.
- Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. 2002. URL <http://mallet.cs.umass.edu/>.

- David Mimno, Hanna Wallach, and Andrew McCallum. Community-based link prediction with text. In *Proceedings of the NIPS 2007 Workshop on Statistical Network Modeling*, 2007.
- M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004. doi:10.1103/PhysRevE.69.026113. URL <http://link.aps.org/doi/10.1103/PhysRevE.69.026113>.
- Evelien Otte and Ronald Rousseau. Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6):441–453, 2002. doi:10.1177/016555150202800601. URL <http://jis.sagepub.com/content/28/6/441.abstract>.
- Tan Pang-Ning, Michael Steinbach, Vipin Kumar, et al. *Introduction to data mining*, chapter 8. Addison-Wesley, 2006.
- Carlos Andre Reis Pinheiro. *Social network analysis in telecommunications*, volume 37. John Wiley & Sons, 2011.
- Howard Rheingold. *The virtual community: Homesteading on the electronic frontier*. MIT press, 1993.
- Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press. ISBN 0-9749039-0-6. URL <http://dl.acm.org/citation.cfm?id=1036843.1036902>.
- Yong Yin and Kazuhiko Yasuda. Similarity coefficient methods applied to the cell formation problem: a taxonomy and review. *International Journal of Production Economics*, 101(2):329–352, 2006.
- Haizheng Zhang, Baojun Qiu, C.L. Giles, Henry C. Foley, and J. Yen. An lda-based community structure discovery approach for large-scale social networks. In *Intelligence and Security Informatics, 2007 IEEE*, pages 200–207, 2007. doi:10.1109/ISI.2007.379553.
- Yang Zhang, Y Wu, and Q Yang. Community discovery in twitter based on user interests. *Journal of Computational Information Systems*, 8(3):991–1000, 2012.
- Ding Zhou, Eren Manavoglu, Jia Li, C Lee Giles, and Hongyuan Zha. Probabilistic models for discovering e-communities. In *Proceedings of the 15th international conference on World Wide Web*, pages 173–182. ACM, 2006.