ASONAM '15, August 25-28, 2015, Paris, France
ACM ISBN 978-1-4503-3854-7/15/08
DOI: http://dx.doi.org/10.1145/2808797.2808840

Beyond Friendships and Followers: The Wikipedia Social Network

Johanna Geiß, Andreas Spitz, Michael Gertz
Institute of Computer Science, Heidelberg University
Im Neuenheimer Feld 348, 69120 Heidelberg, Germany
Email: {johanna.geiss, spitz, gertz}@informatik.uni-heidelberg.de

***Abstract*—Most traditional social networks rely on explicitly given relations between users, their friends and followers. In this paper, we go beyond well structured data repositories and create a person-centric network from unstructured text – the Wikipedia Social Network. To identify persons in Wikipedia, we make use of interwiki links, Wikipedia categories and person related information available in Wikidata. From the co-occurrences of persons on a Wikipedia page we construct a large-scale person-centric network and provide a weighting scheme for the relationship of two persons based on the distances of their mentions within the text. We extract key characteristics of the network such as centrality, clustering coefficient and component sizes for which we find values that are typical for social networks. Using state-of-the-art algorithms for community detection in massive networks, we identify interesting communities and evaluate them against Wikipedia categories. The Wikipedia social network developed this way provides an important source for future social analysis tasks.**

I. Introduction

Looking at the increasing amount of research activities in the area of online social networks in the past few years, it becomes evident that most analyses focus on what we call explicit person-centric networks. Examples include prominent social media sites such as Twitter, Facebook, or LinkedIn, to name but a few. There, users register themselves, provide personal information and establish links to other users based on social relationships such as *friendship*, *follower*, or by simply sending messages to each other. For social analysis tasks, the network that consists of nodes representing persons or users and edges describing a relationship between persons is explicitly given in these instances. Even though such types of networks are rich in terms of node and edge attributes as well as temporal information, it is important to recognize that person-centric networks occur in many other settings as well.

Besides creating social network from surveys, which is a standard approach in sociology, person-centric networks have also been extracted (semi-)automatically from diverse types of publicly accessible semi-structured data repositories. These include email archives, e.g., [1], [2], [3], the blogosphere, e.g., [4], digital libraries from which co-authorship networks are derived, e.g., [5], [6], [7], movie databases [8], and many others. For a more comprehensive overview, see the introductory chapters in [9]. A common challenge in these respective approaches for the extraction of person-centric networks is that the network is latently embedded in the data. Therefore, suitable methods for the extraction of persons and relationships from the raw data are crucial. While many such approaches work fairly well, they are mostly tailored towards a specific type of domain, e.g., authors of papers or actors in movies, and often result in network structures that are relatively small in comparison to major social media sites.

In this paper, we present a framework to extract a large-scale person-centric network structure from the English Wikipedia which contains more than 5.6 million documents.[1] Our main objective is the extraction of a network structure that (1) is large compared to other, more specialized networks, (2) deals with persons and communities that are mostly well-known, and (3) can easily be combined with other data, e.g., DBpedia and in particular Wikidata, the latter being an important component in our approach as detailed below. We base our approach on co-occurrences of person mentions on Wikipedia pages. Here, the hypothesis is that if two persons are mentioned on a Wikipedia page, then there is a common context or theme to which both belong. The closer the two names appear on a page, the more evident is the relationship between the persons.

Rather than relying on linguistic tools for named entity recognition (NER) to identify mentions of persons, we make use of interwiki links to person pages, which can easily be identified by their Wikipedia category. In contrast to NER, this approach has the advantage that no disambiguation is necessary and the precision for the identification of person mentions is optimal. Furthermore, it is an easy task to enrich the created network with additional person information. For this, we use Wikidata as a repository of person names and attributes, which provides a structured data repository that is built on data extracted from Wikipedia, thus realizing a "Wikipedia for data" [10]. Wikidata contains data about roughly 1.2M persons who are represented by a page in the English Wikipedia. It also provides us with unique person ids and further person attributes, the latter being useful for analysis tasks such as the modularity of the resulting person network. The extraction of person mentions based on interwiki links already provides some interesting insights into the link usage and structure for persons in Wikipedia. We develop a measure for weighting an edge between two persons in the network based on the number of co-occurrences in the text of pages and on the distance between the persons mentions. Through this approach and based on an experimentally determined threshold for edge weights, we obtain a person network consisting of roughly 800k persons and 67M edges, which we refer to as the *Wikipedia social network*[2].

[1] Throughout the paper, we refer to the version from January 12, 2015.

[2] The wikipedia social network is available for download at our website: `http://dbs.ifi.uni-heidelberg.de/index.php?id=data`

For this person-centric network, we conduct a number of typical analytical tasks to determine key characteristics such as centrality, clustering coefficient or components sizes and show that the Wikipedia social network exhibits properties that are typical for most social networks. We also identify interesting communities and evaluate them based on the Wikipedia categories that can be associated with members of the communities. In general, we expect that the Wikipedia social network provides an important source for future social network analysis tasks that concentrate on mostly well-known and relevant persons and in cases where one cannot rely on commercial social media site data.

The remainder of the paper is structured as follows. After a review of related work in Section II, we detail the extraction of person names as nodes for our social network in Section III. The construction of the network with a focus on determining (weighted) edges is presented in Section IV, followed by key networks characteristics described in Section V. An evaluation of communities is presented in Section VI before concluding the paper in Section VII.

II. Related Work

Most approaches for extracting person-centric and social networks from semi-structured data such as documents and Web pages concentrate on co-authorship and academic collaboration networks, utilizing rich corpora of bibliographic information. Elmacioglu and Lee [11] did a comprehensive bibliometric study of the DBLP collaboration network. Barabási et al. [5] as well as Huang et al. [12] focus on the evolution of such scientific collaboration networks over time. The well-known system ArnetMiner [6], [7] is another type of comprehensive framework for extracting a collaboration as well as a citation network. In the approach presented by Yang et al. [13], authors, documents, citations and venues are modelled in an integrated framework. Common to the above approaches is that co-authorship relationships can easily be identified from bibliographic sources, because author names mostly co-occur in a standard way and often come with additional metadata such as affiliation or venue as well. Co-authorship networks, however, are by their nature limited to a specific group of people (namely academics). Furthermore, it is a challenge to obtain additional information for the persons in such a network, a problem which is not present in the case of Wikipedia, where additional information is abundant and the set of persons is very diverse, including historic figures.

Another large body of work focusing on the extraction of person-centric networks exists for mail archives and blogs. The work by Diesner at al. [3] is probably one of the most prominent approaches where the communication structure based on the Enron email corpus has been investigated. Other works that focus on email archives and blogs include, for example, [2], [4] and [14]. Analogously to co-authorship, person mentions and communication patterns in email and blogs allow for the simple extraction of networks due to their respective structure. A similar argument holds for communication patterns and person network structures in open software projects [1], [15].

There are a few approaches to extract person-centric network structures from Wikipedia. Maniu et al. [16] study the role and interaction of contributors in Wikipedia, while Massa [17] aims at recovering a network structure from Wikipedia talk pages. Similarly, Sepehri Rad et al. [18] extract collaboration patterns among editors of Wikipedia pages. While all of the above approaches focus on social network aspects in terms of authorship and expertise in editing Wikipedia pages, few approaches aim at discovering social networks from person mentions on Wikipedia pages themselves. Liu et al. [19] investigate the construction of an egocentric network for a given person from Chinese Wikipedia pages using a semantic parser for obtaining relationships. However, they do not construct a comprehensive network of all person mentions of a large set of Wikipedia pages.

While all of the above approaches exploit the co-occurrence of person mentions based on the structure of the data, there are also a few approaches that aim at extracting person-centric networks from general types of documents, in particular Web pages. One of the earliest work is the Referral Web system developed by Kautz et al. [20], [21] to establish "chains" between individuals. Key to the approach is to exploit co-occurrence of names in Web documents as evidence of a relationship between respective persons. However, it is not clear which techniques are used to identify person mentions on a page and thus to what extent the resulting network is comprehensive. Matsuo et al. [22] follow a similar approach, again based on the co-occurrence of names on Web pages, but also include the extraction of different types of relationships between persons. In their approach, a list of persons is explicitly given, and the evaluation is only conducted in the context of academic networks, thus making this another useful albeit domain specific approach.

III. Extracting person mentions

The fundamental step in the creation of the Wikipedia social network is the identification of person mentions within each Wikipedia content page. Wikidata plays a crucial role in this step and provides additional information for further analysis tasks. In the following, we first describe Wikidata, how person related information is extracted from it, and give an overview of the data. Then, we explain how person mentions in Wikipedia are identified in a two step approach and provide results, properties and characteristics for each step.

A. Extracting person information from Wikidata

Wikidata is a free, collaboratively edited, multilingual database by the Wikimedia Foundation that was launched in October 2012 [10]. The intention is to provide one common source of structured information that supports other Wikimedia projects. As of January 26, 2015, Wikidata includes more than 16.8M items, which represent real life topics, concepts, or subjects. Each item is described by a unique identifier, a label, a description, possible aliases and statements that characterize the item in detail.

In a first step, we extract about 2.6M person entries from Wikidata that are classified as an `instance of human`. For most of these persons, Wikidata provides additional information, such as gender, date of birth, date of death, occupation, country of citizenship or links to Wikipedias. 45% (≈ 1.2M) of the person entries have a link to the English Wikipedia and are therefore relevant to our approach. In the following, all statistics and numbers refer to this subset.

TABLE I. TOP OCCUPATIONS OF PERSONS IN THE WIKIDATA SUBSET AND IN THE SET OF REFERENCED PERSONS IN WIKIPEDIA.

Wikidata		references		
rank	%	rank	%	occupation
1	15.48%	2	12.44%	politician
2	14.48%	12	3.16%	association football player
3	7.37%	1	13.90%	actor
4	3.73%	4	8.06%	singer
5	3.29%	15	2.82%	sportsperson
6	3.25%	3	8.57%	author
8	2.93%	5	7.57%	writer

The gender ratio within the data set is unbalanced: 84.3% of persons are male, 15.6% female, and 0.1% have another or unknown gender. 64.8% of the persons have one or more entries for occupation. By far the largest occupational groups are politicians and football players (see Table I for details). $83,75\%$ of the person entries have information on the date of birth or date of death. With this data at hand, we find that Wikidata focuses on persons from the 19th century to the present, most of which are alive today.

B. Recognition of persons in Wikipedia

For our approach, we use the dump of the English Wikipedia from January 12, 2015. It contains about 5.29M content pages.[3] The text content was cleaned from Wikipedia mark-up and split into sentences using the NLTK [23], resulting in about 65M sentences. Using category information from Wikipedia, pages about persons are extracted as follows. We classify a Wikipedia page as a person page when it belongs to at least one of the categories ⟨year⟩ `deaths` or ⟨year⟩ `births`. This approach results in $1,052,898$ pages about persons, which are $178,807$ persons less than we find in Wikidata. The reasons for this difference are 1) the heuristic used for extracting person pages and 2) the difference in the creation date of the dumps, since the Wikidata dump is dated 2 weeks after the Wikipedia dump. During this time period, new pages were added (on average 909 per day [24]), so that we identify links in Wikidata entries to Wikipedia pages that are not present in our Wikipedia dump. With the simple heuristic used, we cannot identify all pages about persons. Unfortunately, Wikipedia contains no specific category for humans. Including more categories such as `Date of birth missing (living people)`, ⟨year⟩`s/`⟨y⟩`th century births`, `Living people`, and `Possibly living people` would likely increase the number of identified Wikipedia person pages. Since Wikidata is a reliable source of person names and the categories are only used as a backup-method when no information is found in Wikidata, we can neglect these shortcomings.

To find references to persons within the set of all Wikipedia pages, we now use a two-step approach:

1) Following so-called *interwiki links* (IWLs)
2) Searching for recognized person names

1) Following interwiki links (IWLs). IWLs are links within a Wikipedia page to another Wikipedia page. An IWL has the form [[*linkTarget*|*coveredText*]], where *coveredText* is optional. In order to determine which IWLs refer to a person, we use the link information in Wikidata and the category information from Wikipedia (see Algorithm 1).

[3]Page titles starting with `File:`, `Category:`, `Wikipedia:`, `Portal:` and `Book:` were omitted.

Algorithm 1 determine person links

```
for each IWL in Wikipedia do:
    linkTarget ← first part of IWL
    if linkTarget == link in Wikidata person entry then
        IWL links to person
    else if linkTarget is a redirect then
        if redirect == link in Wikidata person entry then
            IWL links to person
        end if
    else
        get categories of linkTarget
        if category matches "^\d+(s)? (BC )?births|deaths$"
    then
            IWL links to person
        end if
    end if
end for
```

The English Wikipedia contains about 76.8M IWLs, of which 13.6% (≈ 10.4M) refer to persons. Most of the person links are identified using the link information in Wikidata and only 0.1% were found using Wikipedia categories. We label each person that is referenced in this way with either the corresponding Wikidata id or its Wikipedia page id as a unique identifier. About one third of the content pages in Wikipedia ($1,894,392$ pages) contain IWLs to a person. The person-IWLs refer to $842,484$ different persons, most of which are listed in Wikidata ($99,6\%$). The person that is linked most often in IWLs is Barack Obama, followed by George W. Bush (see Table II).

2) Searching for recognized person names. Next, we refine our results by including references to persons outside of IWLs. Here, we search each page for persons that are already referred to in an IWL on that page. Every page is searched for the *linkTarget* and the *coveredText*. About one third of pages that link to a person page also contain references to persons outside of IWLs. In total, we find $2,695,787$ references outside of IWLs to $273,166$ persons in $631,183$ pages.

Combining the two methods, we find a total of $13,140,069$ mentions of persons in $1,894,392$ Wikipedia articles. $510,309$ of these articles have only one reference (IWL) to a person. The page with the most mentions of persons is `Rosters of the top basketball teams in European club competitions` with $4,694$ mentions of $1,761$ different persons. The article with mentions of the most persons ($2,658$ persons in $2,686$ mentions) is `List of Test cricketers`. The cumulative distribution of links and persons per articles is shown in Figure 1. It shows that the majority of pages have up to three references to persons. The persons that are most commonly referenced overall are Jesus and Napoleon (see Table II).

TABLE II. TOP REFERENCED PERSONS

IWL			all references		
rank	#	# articles	rank	#	person
14	5,209	4,645	1	24,771	Jesus
3	8,918	7,908	2	18,592	Napoleon
1	11,954	10,179	3	16,034	Barack Obama
27	4,255	3,454	4	14,444	Muhammad
4	8,770	8,004	5	14,258	William Shakespeare
5	8,229	7,348	6	14,104	Adolf Hitler
2	10,650	9,312	7	14,072	George W. Bush

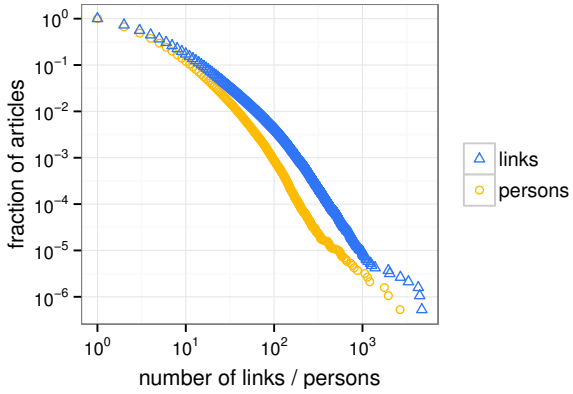

Fig. 1. Cumulative distribution of the number of persons and the numbers of links in the data set plotted against the number of articles.

$83,8\%$ of the IWLs link to male persons and $15,8\%$ to female persons, which approximately reflects the gender ratio in the set of referenced persons as well as in Wikidata and Wikipedia. In Figure 2, the cumulative fractions of birth and death dates of persons who are linked in Wikipedia are shown, limited to persons for which this information is available. This shows that Wikipedia covers mostly contemporary topics and links to persons from recent history. In the set of referenced persons, most persons belong to the occupational groups actor and politician (for details, see Table I).

$839,490$ persons from the Wikidata subset are referenced in the English Wikipedia, meaning that there are $392,215$ persons in Wikidata that have a link to the English Wikipedia (and thereby a page in Wikipedia) but are not referenced in the rest of Wikipedia. These persons might be referenced in an infobox or within a table. For our approach described here, we only extract the content from the continuous text of Wikipedia pages, omitting tables, infoboxes, or captions of images.

IV. Network Construction

To construct the person-centric network, we consider the co-occurrence of references to identified persons as described above. In total, we observe $309,292,211$ co-occurrences between a total of $799,181$ persons. Formally, this provides a bipartite graph $B = (V \cup D, E_B)$, where V is a set of nodes that corresponds to the persons, D a set of nodes that corresponds to the set of Wikipedia documents. For two nodes $v \in V$ and $p \in D$, E_B contains an edge (v, p) if and only if the document p contains person v. To obtain a network of persons, we project this bipartite graph onto the set of persons V. Since a Wikipedia page with k mentions of persons induces $\binom{k}{2}$ edges between persons, the density of the resulting projection is considerable. In the following, we thus describe the scheme used for weighting the resulting edges.

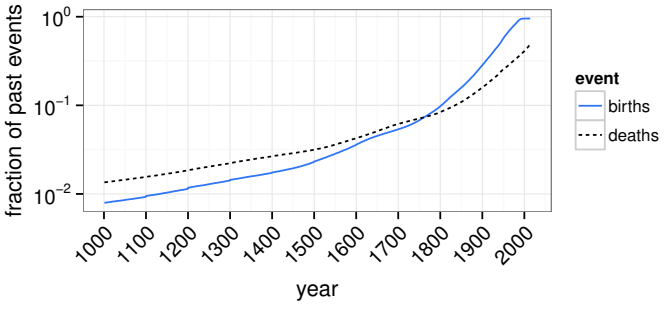

Fig. 2. Cumulative fraction of births (blue) and deaths (dashed) of persons referenced in Wikipedia plotted against the year. The total number of persons for which a known birth or death date exists is used for normalization.

In a first step, we create a multi-graph as projection in which each co-occurrence between two persons induces exactly one edge. In a second step, we then aggregate the edges to obtain a simple graph without multiple edges. Let $M = (V, E_M)$ denote the multi-graph over the set of all persons V. Two persons $v, w \in V$ are connected by an edge $e = (v, w, i) \in E_M$ if there exists an instance i where v and w co-occur in a document. Note that co-occurrence instances need not necessarily correspond to documents, since a document may contain multiple instances of a given person. As a result, E_M contains a distinct edge for each co-occurrence of v and w. Given an instance of a co-occurrence i between persons v and w, we define the distance between those persons $d(v, w, i)$ as the number of sentences that separate the occurrences of v and w in the document that corresponds to instance i, or 0 if they occur in the same sentence. To create weights for edges in M based on this distance, we introduce a weight for the edges of the projection that decays exponentially, i.e. $\varphi : E_M \to \mathbb{R}$, as

$$\varphi(e = (v, w, i)) := \exp\left(-\frac{d(v, w, i)}{2}\right)$$

To aggregate multiple parallel edges into a single edge, we use a cosine similarity of adjacency vectors of nodes in the weighted node-edge incidence matrix of M. This corresponds to a weighted cosine similarity of neighbourhoods for the two incident nodes. Therefore, let n_v denote the neighbourhood of person v in M and $n_v \cap n_w$ the shared neighbourhood of persons v and w in M. Then we obtain a cosine similarity of node-edge incidence vectors based on the decaying distance measure as

$$dicos(v, w) := \frac{\sum_{e \in n_v \cap n_w} \varphi(e)^2}{\sqrt{\sum_{e \in n_v} \varphi(e)^2} \sqrt{\sum_{e \in n_w} \varphi(e)^2}}$$

To generate the weights of edges in a simple projection $G = (V, E)$, where $e = (v, w) \in E$, we define a weight function $\omega : E \to \mathbb{R}$ as $\omega(e) := dicos(v, w)$. As a point of reference, we also consider a weight function $c : E \to \mathbb{N}$ that simply assigns to an edge $e = (v, w)$ the combined number of co-occurrences of v and w over all documents.

V. Properties of the Network

The resulting aggregated network contains $67,583,553$ edges that connect the $799,181$ persons, 98.8% of which are contained in a single giant component. In the following, we therefore describe the process of selecting a threshold for edge weights and creating a sparser network, before we turn to two possible applications for this network, namely network centrality and community detection.

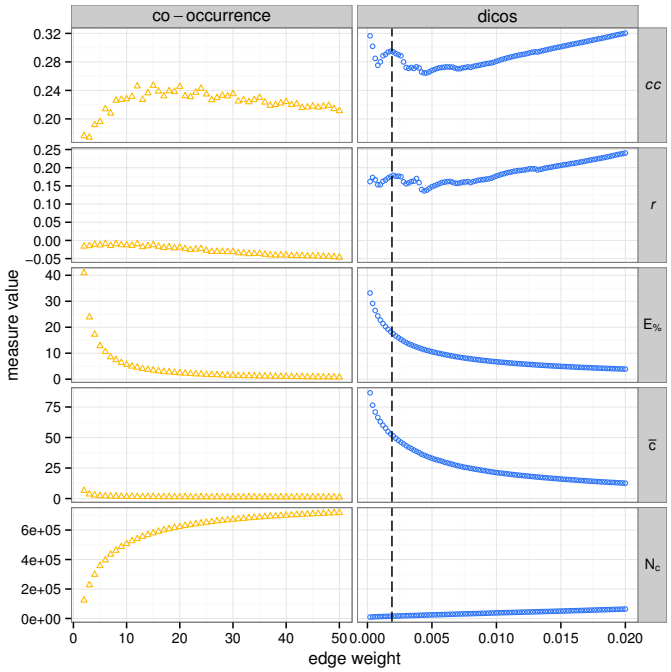

Fig. 3. Structural graph metrics plotted against the minimum threshold of edge weights. Shown are the clustering coefficient cc, the assortativity by degree r, the percentage of remaining edges $E_{\%}$ compared to the entire network, the average component size $\bar{c}$, and the number of connected components N_c. The dashed line denotes a *dicos* threshold of $t_\omega = 0.0019$

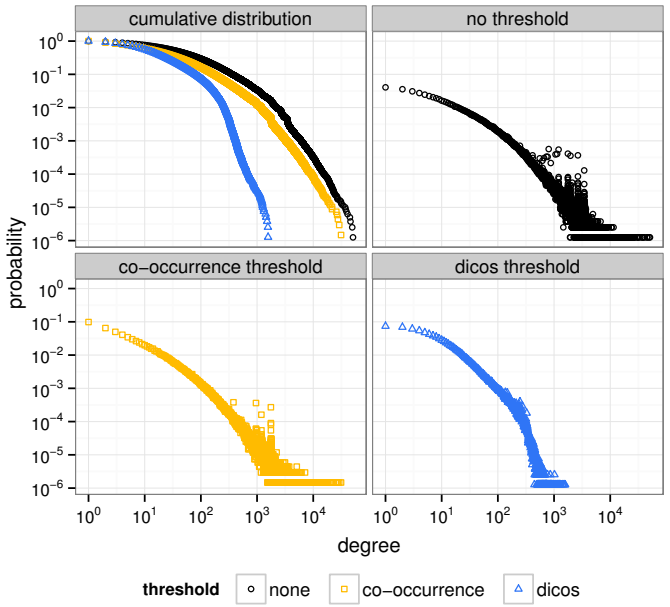

Fig. 4. Complementary cumulative degree distribution of degrees $P[d \geq d_k]$ (top left) and normalized degree distributions $P[d = d_k]$ (bottom and right), where d_k is the respective degree on the horizontal axis. Shown are the degrees for nodes in the entire social network (black circles), the network that results from applying a co-occurrence threshold $t_c = 2$ (orange squares) and the resulting network after using the *dicos* threshold of $t_\omega = 0.0019$ (blue triangles).

A. *Threshold selection*

Since the number of edges in the resulting graph is enormous, finding a threshold for edge weights to reduce the number of edges prior to an analysis is beneficial. To minimize the loss of information, it is advisable to consider structural metrics of the network in this process (for a sociological motivation see Freeman [25], for an application to large scale networks see Serrano et al. [26]). Here, we include in this approach the clustering coefficient, the assortativity by degree and the number and size of components, since we expect a social network to show assortative mixing and consist of dense communities. Path-based graph metrics such as diameter or average path length are also worth considering. However, due to the size of the network, computing the exact values of these metrics for each possible threshold is not feasible [27] and using an approximation would defeat the purpose of finding a proper threshold.

In Figure 3, we show the result of such an exploration as a plot of graph measures against the threshold used for edge removal (all edges with a lower value were removed) for both the *dicos* weight and the co-occurrence weight. Note that we only show the results for a small subset of possible thresholds, since the number of removed edges is too extreme for higher values. However, all shown measures progress monotonously beyond the shown values. Here, one observes that co-occurrence is not well suited as a weight for the purpose of thresholding. Even if we only remove edges with a co-occurrence of 1, the graph decomposes almost completely into a large number of small connected components. The clustering coefficient after this step is relatively low and only slowly rises as the network decomposes further into negligibly small components. Applying a threshold to the *dicos* weight is more promising on the other hand. One can observe a clear peak in the clustering coefficient and assortativity score at a value of $\omega = 0.0019$. The average size of connected components is still fairly high at this point and the percentage of remaining edges is still 20%. Therefore, we select $t_\omega = 0.0019$ as threshold and remove edges with a lower weight. For the purpose of comparison, we also consider the entire network as well as a network in which edges with a co-occurrence weight of a count less than $t_c = 2$ are removed.

In Figure 4, the degree distributions for the entire Wikipedia social network and the two networks to which we applied a threshold of t_ω and t_c, respectively, are shown. The distributions have a distinct long tail as one would expect from a social network. The distribution of the entire network also includes atypical outliers that are created by Wikipedia pages containing large lists of persons, most of which only occur on this one page. Since some of these persons have a slightly higher degree than others, one can observe vertical streaks for each such list. Interestingly, the thresholding by *dicos* weight almost completely removes these trivial cliques from the network.

To assess whether the extracted network can serve as a proxy for social networks, we also consider the temporal properties of edges. In Figure 5, we show the distribution of differences in birth and death dates of adjacent nodes. It shows that the probability drops sharply for time spans that are greater than the average human lifespan. We also find that the network with a *dicos* threshold emphasizes short time spans even more strongly than the other two networks. If we consider the modularity by gender [28], we find that the network shows

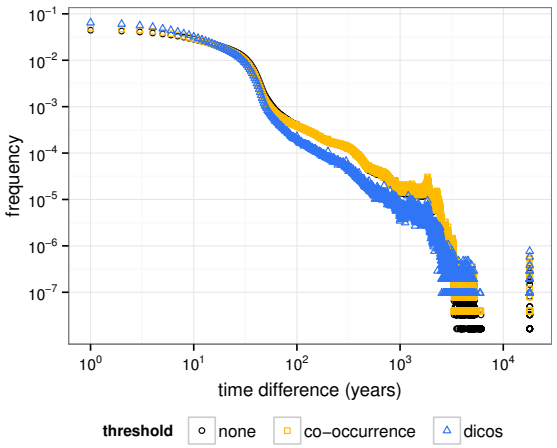

Fig. 5. Distribution of temporal distances between connected persons in the networks. For each pair of adjacent nodes, the pairwise minimum of their respective birth and death dates is considered to account for cases where only one of these dates is known.

TABLE III. THE 20 HIGHEST RANKED PERSONS IN THE NETWORK WITH A *dicos* THRESHOLD OF $t_\omega = 0.0019$ ACCORDING TO PAGERANK.

rank	degree	gender	birth	death	label
1	1561	m	1961		Barack Obama
2	1449	m	1920	2005	John Paul II
3	1419	m	1946		George W. Bush
4	1508	m	1889	1945	Adolf Hitler
5	1249	m	1946		Bill Clinton
6	1232	m	1882	1945	Franklin D. Roosevelt
7	1572	m	1769	1821	Napoleon
8	1141	m	1927		Benedict XVI
9	1217	f	1926		Elizabeth II
10	1130	m	1911	2004	Ronald Reagan
11	1317	f	1819	1901	Queen Victoria
12	1154	m	1809	1865	Abraham Lincoln
13	1197	m	571	632	Muhammad
14	1242	m	1600	1649	Charles I of England
15	947	m	1890	1969	Dwight D. Eisenhower
16	1321	f	1533	1603	Elizabeth I of England
17	981	m	1913	1994	Richard Nixon
18	1140	m	1732	1799	George Washington
19	996	m	1858	1919	Theodore Roosevelt
20	1056	m	1879	1953	Joseph Stalin

no assortative mixing with $Q = 0.078$, despite the disparity in gender frequencies. These two observation indicate that the network models connections between persons in Wikipedia in a similar way as a directly observed social network.

B. Network Centrality

The question of network centrality is a well researched aspect of social network analysis. It is beyond the scope of this article to perform an in-depth analysis of different types of centralities in the Wikipedia social network. However, due to the direct interpretability, centrality is a good measure to assess the consistency of the created network Therefore, we use the well established PageRank centrality [29], which lends itself naturally to centrality analysis in the context of social networks with a large number of less well-known persons, as it incorporates a propagation of importance. In Table III, we show the highest ranked persons according to PageRank centrality for the network with *dicos* weights and a threshold of $t_\omega = 0.0019$. There is a correlation to the list of top referenced persons (see Table II), which is not unexpected due to the similarity between PageRank in a first iteration and degree centrality. While the degree plays a role in the ranking, there clearly are additional factors of influence at work. The birth years of the highly ranked persons provide evidence that the list is fairly balanced between living and deceased persons. This does not meet the expectation that PageRank would rank persons more highly whose period of activity lies further in the past, which is a common issue with PageRank. Overall, the results match our intuition about centrality in a network in which influential persons should be ranked highly.

C. Communities

The detection of communities is a central problem, not just with regard to social networks, but also in many other fields of network analysis. As a result, there are many different algorithms and approaches to the problem that vary with regard to their efficiency, applicability to certain networks and even the definition of community itself. For a recent overview, see the review and evaluation by Harenberg et al. [30]. In the case of the Wikipedia social network, a key limiting factor is the size of the network, which excludes the traditional algorithms for community detection as well as those based on clustering techniques, simply due to their time complexity. Furthermore, given the nature of the network, an algorithm that detects overlapping communities is a reasonable choice. Based on the results of the evaluation performed by Harenberg et al., we select the stabilized label propagation algorithm (SLPA) as community detection algorithm, since it is the best performing algorithm that meets the requirements [31]. The algorithm assigns a unique label to each node in the network and, in each subsequent step, adds a label of adjacent nodes by majority vote. Since nodes may accumulate multiple labels this way, the resulting clustering is soft and allows for overlapping communities, with the probability of community membership depending on the distribution of labels for each node. To obtain a hard clustering, a node can be assigned to the most probable community. To reduce the number of communities,

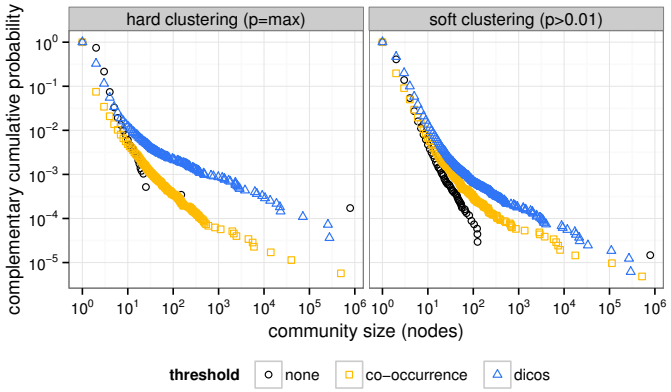

Fig. 6. Distribution of community sizes for the entire network and the two networks that are obtained by applying a co-occurrence and *dicos* threshold. Shown are the distributions for a hard clustering, where each node is assigned to the community with maximum probability value, and a soft clustering in which each node may belong to multiple communities. In the soft clustering, a probability threshold of 0.01 was used to prevent each node from being assigned to its own community.

it is advisable to use a threshold for the probabilities even if a soft clustering is needed, since the number of communities would otherwise equal the number of nodes by construction.

Our implementation of SLPA is used for all three constructed networks, i.e., the network with no thresholds and the two thresholded networks by co-occurrence and *dicos*. For each network, labels are propagated over $T = 100$ iterations. In Figure 6, we show the distribution of sizes for the detected communities. For a hard clustering, we find $4,292$ communities in the network with no threshold, $3,584$ communities in the co-occurrence thresholded network and $8,193$ if we use the *dicos* threshold. It is clear that the smaller number of identified communities in the entire network results from the high density that leads to one very large community. Both in the cases of hard and soft communities, the overall number of medium sized communities is highest for the *dicos* thresholded network.

VI. Community Evaluation

In order to evaluate the quality of the identified communities, we need a ground truth for comparison. Using Wikipedia categories for this task is possible, yet not as straightforward as expected due to the large variance of topics and sizes of categories in Wikipedia. The $842,484$ identified persons are listed in $841,158$ categories. Most persons are listed in more than one category, e.g., Winston Churchill has 97 categories. $5,282$ persons are members of just one category. The largest category in this given subset is `living people` with $438,500$ members. Therefore, we use an approach that selects in each step the category that fits best.

A. Evaluation strategy and measures

For each identified community, all categories of all its members are retrieved from the Wikipedia dump. In the next step, each community is compared to all such categories by calculating Precision, Recall and F-score. For each community, the maximum F-score is stored and used to calculate the average F-score over all communities at the end.

$$P = \frac{|community \cap category|}{|community|}$$

$$R = \frac{|community \cap category|}{|category|}$$

$$F = \frac{2PR}{P+R}$$

We evaluate two different sets of communities with three different threshold settings each. The first set includes all communities with more than 10 and less than 500 members, since these correspond to sizes of social groups one would naturally expect. The second set includes all communities that have more than one member. For each set, communities are extracted by using three different threshold settings as described in Section V:

none	no threshold was applied
t_c	co-occurrence threshold of 2
t_ω	distance cosine threshold of 0.0019

TABLE IV. F-score, Precision and Recall for different sets of communities

communities	threshold	# comms	# persons	F	P	R
subset	none	90	1,562	**0.4612**	**0.6138**	**0.4341**
10<n<500	t_c	301	10,683	0.4105	0.5583	0.4078
	t_ω	713	24,315	0.3889	0.4785	0.4238
all	none	4,292	798,777	0.2883	**0.6223**	0.2316
	t_c	3,584	677,880	0.2923	0.6002	0.2467
	t_ω	8,193	788,279	**0.2954**	0.5811	**0.2535**

For this evaluation approach, we use a hard clustering, meaning that each person belongs to only one community. As mentioned in Section V, the number of communities and persons vary between the threshold settings. The results are shown in Table IV.

B. Discussion of results

The quality of communities in comparison to Wikipedia categories depends on the thresholding. If all communities with more than one member are considered, using the *dicos* threshold results in the best communities. This shows that the loss of information is minimized by using the *dicos* threshold (only 1,3% of the person nodes were removed as isolated nodes), while the relevant edges remain in the network. In the range of 10 to 500 members, applying no threshold works best. However, closer inspection of the relevant data reveals that only very few communities and persons are identified in this network. While the *dicos* threshold results in a 16% lower F-score, the recall stays at 0.42 and we find eight times more communities containing 15 times more persons in total.

Another question that arises is the quality of the ground truth. There is no set of clear-cut rules that determine when to place a Wikipedia page into a given category. Some rules, guidelines and policies are given, but these are continuously refined and are subject to change[4]. Even if there were specific guidelines or policies, they are open to interpretation. In a talk about categorization[5] for example, there is a discussion about whether to put a page about a website run by a person from Bristol in the category `People from Bristol`. Further examples include a cartoon series with a robot character being placed in `Robots in fiction`. This shows that category assignment is prone to inconsistency due the vast number of editors and that many categories have no semantic background. A number of categories are based only on location or time. For example, pages about persons in the categories `1976 births`, `People from Massapequa Park, New York` or `Road accident deaths in South Africa` have only the year of birth, the place of origin or mode of death in common, but might not be related at all. Another problem in using Wikipedia categories as ground truth is the hierarchical organization of the categories. To which level in the hierarchy should we compare our communities? A guideline for categorization in Wikipedia is to place a page in the most specific subcategory and not in the parent category as well.

On the other hand, in the Wikipedia social network, there are communities that are semantically very well related. For example, we are able to identify the complete crew of the

[4]see http://en.wikipedia.org/wiki/Wikipedia:FAQ/Categorization#State_of_the_Category_feature

[5]http://en.wikipedia.org/wiki/Wikipedia_talk:Categorization

men's eight of the 1962 British Empire and Commonwealth Games, the members of the coxless four in the 1936 Summer Olympics or all members of the Hungarian heavy metal band Pokolgép. Wikipedia contains no specific categories for these types of groups, yet they are valid communities. In this evaluation, a person was assigned to only the most probable community, which also leads to very large communities, especially in dense networks. It is difficult to meaningfully evaluate them. They can be compared to large categories like `Living People`, but will never achieve high precision or recall, since the people in these categories are not related in a meaningful way. The purpose of our approach is thus not to find communities that resemble Wikipedia categories but to gain a tool that is useful in other tasks such as person name disambiguation.

VII. Conclusion and Ongoing Work

In this paper, we presented a framework for extracting a person-centric network from the English Wikipedia. The combination of interwiki links, Wikidata and Wikipedia categories proved to be a valuable tool for determining person mentions, a task for which Wikidata provided crucial information. The Wikipedia social network created from the co-occurrences of identified persons exhibits the typical properties of social networks, such as the high clustering coefficient and assortativity. We found that the difference in birth and death dates corresponds to our expectations for a real-world social network. Furthermore, centrality measures result in natural and reasonable rankings. We introduced a method of weighting the relationship between two persons in the network based on the distance of their respective occurrences within the text. In an evaluation against Wikipedia categories, we showed that a threshold based on our weighting scheme provides the best trade-off between edge reduction and information loss and even serves to increase the number of identified communities.

Currently, we work on further improving the coverage of identified person mentions. For this, we use the Stanford Named Entity Recognizer to find person mentions even outside of interwiki links. In order to establish relationships between persons, the communities can be refined by taking occupation, country of citizenship or biographical data into account. The Wikipedia social network can then be applied to other tasks such as person name disambiguation or serve as a proxy for real-world social networks on a scale where such information would not otherwise be available.

References

[1] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan, "Mining email social networks," in *MSR '06*, 2006, pp. 137–143.

[2] A. Culotta, R. Bekkerman, and A. McCallum, "Extracting social networks and contact information from email and the web," in *CEAS '04*, 2004.

[3] J. Diesner, T. L. Frantz, and K. M. Carley, "Communication networks from the Enron email corpus "It's always about the people. Enron is no different"," *Comput. Math. Organ. Theory*, vol. 11, no. 3, pp. 201–228, 2005.

[4] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," *World Wide Web*, vol. 8, no. 2, pp. 159–178, 2005.

[5] A. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A*, vol. 311, no. 34, pp. 590 – 614, 2002.

[6] J. Tang, D. Zhang, and L. Yao, "Social network extraction of academic researchers," in *ICDM '07*, 2007, pp. 292–301.

[7] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *KDD '08*. ACM, 2008, pp. 990–998.

[8] M. E. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications," *Physical review E*, vol. 64, no. 2, p. 026118, 2001.

[9] M. Newman, *Networks: an introduction*. Oxford University Press, 2010.

[10] D. Vrandečić and M. Krötzsch, "Wikidata: A Free Collaborative Knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[11] E. Elmacioglu and D. Lee, "On six degrees of separation in DBLP-DB and more," *SIGMOD Record*, vol. 34, no. 2, pp. 33–40, 2005.

[12] J. Huang, Z. Zhuang, J. Li, and C. L. Giles, "Collaboration over time: Characterizing and modeling network evolution," in *WSDM '08*, 2008, pp. 107–116.

[13] Z. Yang, L. Hong, and B. D. Davison, "Academic network analysis: A joint topic modeling approach," in *ASONAM '13*, 2013, pp. 324–333.

[14] S. Tavassoli, M. Moessner, and K. A. Zweig, "Constructing social networks from semi-structured chat-log data," in *ASONAM '14*, 2014, pp. 146–149.

[15] M. Y. Allaho and W.-C. Lee, "Analyzing the social ties and structure of contributors in open source software community," in *ASONAM '13*. ACM, 2013, pp. 56–60.

[16] S. Maniu, B. Cautis, and T. Abdessalem, "Building a signed network from interactions in wikipedia," in *DBSocial '11*, 2011, pp. 19–24.

[17] P. Massa, "Social networks of wikipedia," in *HT '11*, 2011, pp. 221–230.

[18] H. Sepehri Rad, A. Makazhanov, D. Rafiei, and D. Barbosa, "Leveraging editor collaboration patterns in wikipedia," in *HT '12*. New York, NY, USA: ACM, 2012, pp. 13–22.

[19] M. Liu, Y. Xiao, C. Lei, and X. Zhou, "Social relation extraction based on chinese wikipedia articles," in *CLSW '12*, 2013, pp. 94–101.

[20] H. Kautz, B. Selman, and M. Shah, "Referral web: Combining social networks and collaborative filtering," *Commun. ACM*, vol. 40, no. 3, pp. 63–65, 1997.

[21] ——, "The hidden web," *AI magazine*, vol. 18, no. 2, p. 27, 1997.

[22] Y. Matsuo, J. Mori, M. Hamasaki, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka, "POLYPHONET: an advanced social network extraction system from the web," *J. Web Sem.*, vol. 5, no. 4, pp. 262–278, 2007.

[23] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.

[24] Stats.wikimedia.org. (12015) Wikipedia statistics - tables - new articles per day. [Online]. Available: http://stats.wikimedia.org/EN/TablesArticlesNewPerDay.htm.

[25] L. C. Freeman, "The sociological concept of "group": An empirical test of two models," *American journal of sociology*, pp. 152–166, 1992.

[26] M. Á. Serrano, M. Boguñá, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proc. Natl. Acad. Sci.*, vol. 106, no. 16, pp. 6483–6488, 2009.

[27] T. M. Chan, "All-pairs shortest paths for unweighted undirected graphs in o(mn) time," *ACM T. Algorithms*, vol. 8, no. 4, p. 34, 2012.

[28] M. E. Newman, "Mixing patterns in networks," *Physical Review E*, vol. 67, no. 2, p. 026126, 2003.

[29] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[30] S. Harenberg, G. Bello, L. Gjeltema, S. Ranshous, J. Harlalka, R. Seay, K. Padmanabhan, and N. Samatova, "Community detection in large-scale networks: a survey and empirical evaluation," *Wiley Interdiscip Rev Comput Stat*, vol. 6, no. 6, pp. 426–439, 2014.

[31] J. Xie and B. K. Szymanski, "Towards linear time overlapping community detection in social networks," in *Advances in Knowledge Discovery and Data Mining*, 2012, pp. 25–36.