The Role of Structural Information for Designing Navigational User Interfaces

Dimitar Dimitrov
GESIS
Cologne, Germany
dimitar.dimitrov@gesis.org

Philipp Singer
GESIS
Cologne, Germany
philipp.singer@gesis.org

Denis Helic
Graz University of Technology
Graz, Austria
dhelic@tugraz.at

Markus Strohmaier
GESIS and University of Koblenz-Landau
Cologne, Germany
strohmaier@uni-koblenz.de

ABSTRACT

Today, a variety of user interfaces exists for navigating information spaces, including, for example, tag clouds, breadcrumbs, subcategories and others. However, such navigational user interfaces are only useful to the extent that they expose the underlying topology—or network structure—of the information space. Yet, little is known about which topological clues should be integrated in navigational user interfaces. In detail, the aim of this paper is to identify what kind of and how much topological information needs to be included in user interfaces to facilitate efficient navigation. We model navigation as a variation of a decentralized search process with partial information and study its sensitivity to the quality and amount of the structural information used for navigation. We experiment with two strategies for node selection (quality of structural information provided to the user) and different amount of information (amount of structural information provided to the user). Our experiments on four datasets from different domains show that efficient navigation depends on the kind of structural information utilized. Additionally, node properties differ in their quality for augmenting navigation and intelligent pre-selection of which nodes to present in the interface to the user can improve navigational efficiency. This suggests that only a limited amount of high quality structural information needs to be exposed through the navigational user interface.

Categories and Subject Descriptors: H.5.3 **[Information Interfaces and Presentation]**: Group and Organization Interfaces—*Web-based interaction* H.5.4 **[Information Interfaces and Presentation]**: Hypertext/ Hypermedia—*Navigation*

Keywords: Navigation; Decentralized Search; Structure; Networks; User Interfaces

HT'15, September 1–4, 2015, Guzelyurt, TRNC, Cyprus.

DOI: http://dx.doi.org/10.1145/2700171.2791025.

1. INTRODUCTION

With the increasing amount of information made available to people on the Web every day, it has become increasingly difficult to build information systems that can be navigated in an efficient way. Information systems that deliver strong intuition about the choices made available to their users through the interfaces are efficient at guiding the user to the needed piece of information. Thus, they are considered good at supporting activities such as *navigation* or *browsing*. In order to improve navigability, new interfaces—e.g., tag clouds, breadcrumbs, subcategories—have been introduced. In Figure 1, we see an example of a tag cloud. Besides other aspects of tag cloud design [27], tag clouds—as well as all other kinds of user interfaces—are only useful for augmenting navigation to the extent to which they are able to expose the underlying structure of the information space [10]. Yet, little is known about what kind of and how many topological clues should be integrated in navigational user interfaces.

Problem. Consequently, in this paper, we want to study the problem of properly exposing the topological structure of the information space through an interface. This problem has two dimensions: (i) Which are the important structural properties that contribute to properly exposing the hidden structure of the information space and (ii) how much should we know about them in order to navigate efficiently? Knowing which nodes in a network are important and how to identify them is crucial for navigation. Such knowledge could reduce the amount and nature of information needed for improving the users' understanding about the information space resulting in better navigational efficiency. Subsequently, we next derive and discuss the two main research questions that we want to tackle in this article.

Research questions. (i) What kind of and (ii) how much structural information is needed for efficient navigation? Regarding the first research question, we are specifically interested in deriving important structural properties of the information space that should be exposed through an interface in order to properly guide users' navigation. Related work [2] has suggested that the degree—as a proxy of a node's popularity—is a very good navigational feature in networks with a power law degree distribution. Yet, little is known about the effect of the clustering coefficient as a navigational feature on the efficiency of navigation. The clustering coefficient may be feasible as navigational feature due to its importance for the emergence of the small world property of a network. Small world networks are

Figure 1: **A tag cloud enabling navigation from The Rolling Stones page on last.fm.** Exemplary user interface used for navigation in many online information systems. The tag clouds among other web interfaces are useful to the extent that they expose the underlying structure of the information space. Identifying the most important tags *from a navigational perspective* is crucial for providing efficient support.

known to be particularly navigable[18, 31]. In this paper, we investigate whether nodes with a specific clustering coefficient have an impact on navigation and we study how the clustering coefficient can be used to identify them. Furthermore, regarding the second research question, we are interested in determining the amount of structural information needed for navigation and if this depends on the quality of the structural information.

Approach and methods. We approach the research questions by analyzing the structural properties of four different networks. Initially, we take a look at their shortest path distance, degree and clustering coefficient distributions, and classify them by their expected navigability according to [4]. To model navigation, we use the message-passing algorithm *decentralized search* which is inspired by the small world experiment by Stanley Milgram [22]. Several versions of the algorithm can be found in literature [21, 18, 19, 30, 2, 1]. Decentralized search has already been demonstrated to be useful for modeling navigation in information networks [13]. For studying which and how much information is needed for efficiently navigating a network, we utilize an adaption of the algorithm which we call *partially informed* decentralized search. The partially informed decentralized search models a user who is limited in her exposure to the structure of the information space and thus, has just a weak or limited understanding of the topology of the information space. We study two strategies for selecting important nodes with regard to their popularity and clustering coefficient. With both strategies, the algorithm navigates by popularity. With simulations, we compare the partially informed decentralized search with the random search and the fully informed decentralized search. In our setting, random search corresponds to a user who is clicking at random and has no intuition. We also make a comparison between the two strategies for node selection to test the importance of the exposure of the user to the underlying structure of the information.

Findings and contributions. The most prominent finding is the surprisingly small amount of structural information needed for efficient navigation and the supportive properties of the clustering coefficient for identifying nodes important for navigation. By and large, our findings suggest that only a limited amount of high quality structural information needs to be exposed through the navigational user interface. Additionally, we empirically demonstrate the sensitivity of decentralized search as a navigational model on the kind of structural information utilized. The navigational performance of decentralized search appears to depend on the amount of high quality structural information provided.

Structure. The rest of this paper is organized as follows. After discussing related work in Section 2, we present an adaptation of decentralized search and two strategies for selecting nodes with high structural importance used in the experimental setup in Section 3. In Section 4, we give detailed overview of the used datasets. In Section 5, we present our results and formulate our findings. Next, Section 6 discusses the findings and their implications for the design of navigational user interfaces. Finally, we conclude the paper and provide some directions for future work in Section 7.

2. RELATED WORK

The decentralized search algorithm is inspired by research conducted in the 1970s by Stanley Milgram who studied the structure of the American society and conducted the famous *small world experiment* [22]. For this experiment, Milgram asked randomly selected people from Nebraska to forward a packet to a stock broker in Boston. If participants did not know the target personally, they were asked to forward the packet to personal contact that they thought might know the target better. These persons then should repeat this process. Even though there were quite some restrictions, the experiment showed that the average chain length of letter trails that reached the target was around six.

Motivated by this small world experiment, researchers [21, 18, 19, 30, 2, 1] have developed the so-called *decentralized search algorithm* that tries to find a path between a *start node* and a *target node* in a network by passing a message from a node to one of its immediate neighbors also called *candidate nodes*. What information is available and how it is used for selecting one of the candidate nodes is decisive for the success of the search. For a detailed description of the decentralized search algorithm, please refer to Section 3.1. Next, we delve into related work and discuss navigation using homophily (Section 2.1), navigation using popularity (Section 2.2), models for user navigation (Section 2.3) and the role of clustering for navigation (Section 2.4).

2.1 Navigation Using Homophily

There are different models based on node similarity or homophily for generating small world networks in which decentralized search is very effective. The two main models are *grid-based* and *hierarchy-based*. The first *grid-based model* was proposed by Watts and Strogatz in [31]. This model places nodes on a two-dimensional grid in a way that nodes with high similarity have small grid distance. In order to assure the emergence of the small world property, the model puts long links between the nodes that are similar, but still locally far away on the grid. This model was improved by Kleinberg in [19, 18] where he concentrated on the length of the long links. He showed that efficient search is only possible for certain values of the

clustering exponent of the model which is responsible for placing the long link connections between the nodes.

The *hierarchical model* was proposed independently by Kleinberg [20] and by Watts et al. [30]; these models are also generative. In hierarchical models, similar nodes are placed near to each other in a hierarchy. The probability of two nodes being connected in the hierarchical model not only decreases with their hierarchical distance but also it decreases exponentially. Another generative model was proposed by Boguña et al. in [4] where they assumed that nodes form a *hidden metric space*. The topology of the metric space determines the distance between the nodes in the metric space and models the probability of a link between them in the generated network. The model also possesses a parameter that is responsible for the clustering in the network. This clustering parameter, like the clustering exponent in Kleinberg's grid model, is also responsible for expressing the homophily of the nodes in the network. The main limitation of these models is the global information about the node's position on the grid or in the hierarchy.

2.2 Navigation Using Popularity

Since estimating similarity between nodes is not easy, Adamic et al. [2] concentrated on the degree of nodes. They proposed an algorithm for efficient search in power law networks which makes use of the power law degree distribution to support the node selection. The algorithm keeps track of a node's identity and uses information about the node's degree and the node's neighbors' degree. The biggest difference to the models elaborated in Section 2.1 is the absence of global information about the target node and its position in the network. Adamic et al. showed that degree-based navigation works fairly well in power law degree distributed networks in comparison to Poisson degree distributed networks. Additionally, in power law degree distributed networks, random walks tend to select high degree nodes and achieve good results in those kinds of networks.

2.3 Model for User Navigation

Decentralized search has a long tradition as a model for user navigation in different types of networks. In [13], Helic et al. showed that decentralized search can be used to model user navigation in information networks. The differences and the similarities between the click traces produced by decentralized search with hierarchical background knowledge and actual user navigation were studied by Trattner et al. [29]. Research on the navigational efficiency of different types (broad and narrow) of hierarchical background knowledge conducted by the authors showed that both types are useful. However, broader hierarchies performed better under the limitations introduced by the user interface [11].

2.4 The Role of Clustering

In [31], Watts and Strogatz used the characteristic path length and the clustering coefficient to define the class of navigable networks. The characteristic path length is the averaged shortest path length over all nodes in the network. The clustering coefficient can be interpreted as the probability of a link to exist between two randomly picked neighbors of a node [23]. In a network $G = (V, E)$, where V is a set of nodes and E is a set of edges, $E \subseteq V \times V$. Let $N(u)$ be the neighborhood of the node u and d_u the degree of the node u. The local clustering coefficient $C(u)$ is then defined as the fraction of pairs of neighbors of the node u that are themselves neighbors:

$$C(u) = \frac{|e_{vw} \in E : v, w \in N(u)|}{d_u(d_u - 1)/2}. \quad (1)$$

An alternative definition of the class of navigable networks was given by Boguña et al. [4] who showed how the navigability of a network depends on its degree distribution and its clustering coefficient. In the models described in Section 2.1 and Section 2.2, the clustering exponent plays an important role for the emergence of the small world networks and it is crucial for navigation.

In [16, 17], the authors studied the impact of the clustering exponent on the navigability of a network, i.e., they showed for different network sizes how the change of the clustering exponent affects the effectivity and the efficiency of four different decentralized search versions. In the next Section 3, we will present an adaptation of decentralized search—partially informed decentralized search—and we will use the degree distribution and clustering coefficient of the networks to identify the nodes for which the partially informed decentralized search will be able to make an informed decision.

3. METHODOLOGY

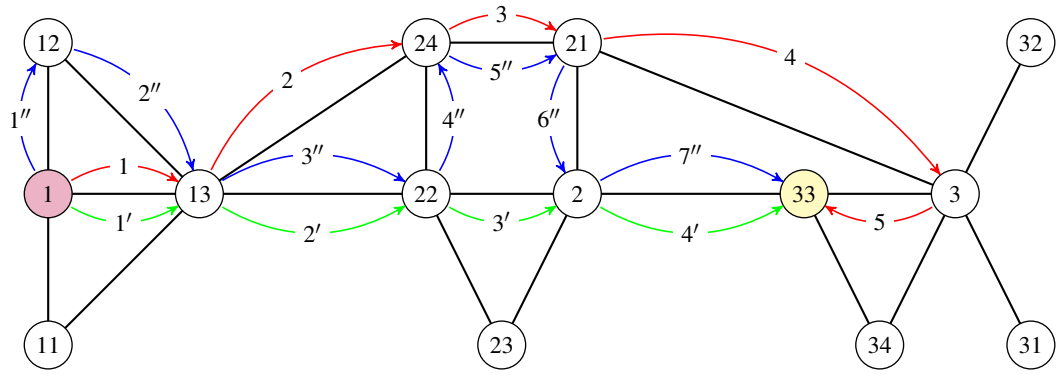
Decentralized search is an established model for navigation. Our goal is to estimate the amount and type of structural information that allows efficient navigation. To this end, we extend the decentralized search algorithm in a way that allows us to simulate navigation with limited amount and different kinds of structural information. By doing so, we can tackle the research questions posed in Section 1. Next, we describe (partially informed) decentralized search in Section 3.1 before we discuss strategies for node selection in Section 3.2 and conduct our experiments in Section 3.3.

3.1 Decentralized Search

In Figure 2, we see an example of both a *fully informed* as well as a *partially informed decentralized search* in a network. The goal is to find the path between node 1 (purple) and node 33 (yellow). The fully informed version of decentralized search uses the degree information as shown in the first row of the table presented in Figure 2 and navigates greedy by degree. Let I be an *informed set* of nodes for which the algorithm can take an informed decision regarding the degree of the candidate nodes. In the case of fully informed search $I = V$, this means that the algorithm possesses the degree information about all nodes in the network. This allows it to rank all candidate nodes by their degree and to select the node with the highest degree. The green arrows show how navigation proceeds for this version of the algorithm. The red arrows show a path produced by the partially informed version of decentralized search. In this version, we only have a fraction of the popularity information as shown in the second row of the table in Figure 2 and the informed set I is a proper subset of V. The partially informed decentralized search ranks the nodes by their degree and selects the node with the highest one only if the set of candidate nodes C contains nodes whose popularity value is available in $I \cap C \neq \emptyset$; otherwise, it picks one node at random. In both versions of the algorithm, we avoid already visited nodes and we terminate the search if the target node is in the set of candidate nodes. For completeness, Figure 2 also highlights an example path of an *uninformed random walker* (blue arrows) that simply picks adjacent nodes at random for navigating.

With the partially informed version of decentralized search, we can estimate the amount of information really needed for navigation in a network. By varying the fraction of the nodes where the popularity is available, we can derive the sensitivity of the algorithm to the amount of popularity information. Thus, this allows us to study our research questions at interest regarding what kind of and how much structural information is necessary for efficient navigation.

Using the methodological concepts explained, we conduct our experiments in Section 3.3. We focus on using the degree of the candidate nodes to model the popularity of nodes. Degree corre-



Popularity/Node	1	11	12	13	2	21	22	23	24	3	31	32	33	34
Degree (fully informed, $I = V$)	3	2	2	5	4	3	4	2	3	5	1	1	3	2
Degree (partially informed, $I \subset V$)	-	-	-	5	-	-	-	-	3	5	-	-	-	-

Figure 2: **Different versions of decentralized search.** *Green:* The arrows show the path produced by a *fully informed decentralized search*. *Red:* The arrows show the path produced by a *partially informed decentralized search*. *Blue:* The arrows represent the path produced by a *uninformed random walker*. The table shows the information provided to the algorithm for selecting the next step. The first row of the bottom table contains the popularity scores of all nodes $I = V$ provided to the fully informed decentralized search and the second row contains only a small portion of all popularity scores $I \subset V$. Fully informed and partially informed decentralized search apply greedy neighbor selection. The partially informed search selects a random node when no information is available. Although finding the shortest path between the nodes 1 (purple node) and 33 (yellow node) is possible with both versions of decentralized search, in general this is not the case because the algorithm can take an informed decision only on the local level.

sponds to the number of links attached to the node [23] and it is a local metric:

$$d_u = \sum_{v \in V} a_{uv} \tag{2}$$

Thus, when we speak about fully and partially informed decentralized search, we speak about fully and partially informed on a local level. If the algorithm was informed on the global level—in other words, if we possessed the adjacency matrix A of the network G—we would be able to calculate the shortest path, which is highly unlikely for real user navigation in large information networks on the web.

3.2 Strategies for Node Selection

In the following, we define two strategies for selecting structurally important nodes: the popularity strategy and the clustering strategy. The nodes selected by these two strategies are elements of the informed set of nodes for which the partially informed decentralized search is going to possess the information about their popularity (i.e., degree) in the network. With these strategies, we can study how the kind of structural information affects navigation.

Popularity Strategy. We sort the nodes by popularity in descending order and take just the top k% of the sorted list. For these nodes, the algorithm will make an informed decision regarding the popularity of the nodes. The idea behind the popularity strategy for node selection is the same as the idea to navigate by popularity, namely highly popular nodes are very well connected. Selecting a highly popular node increases the probability of finding the target node under the nodes' neighbors.

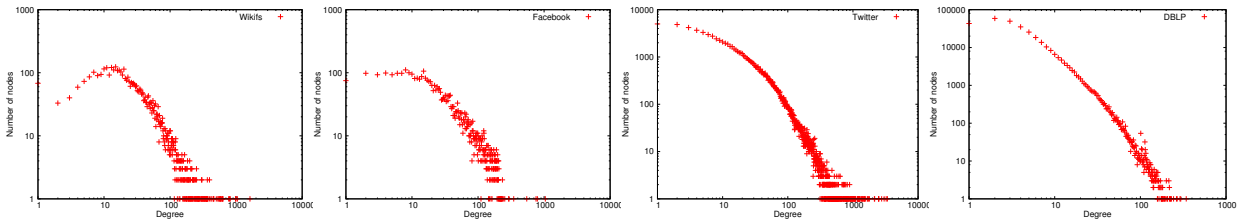
Clustering Strategy. We sort the nodes by clustering coefficient in ascending order and take just the top k% of the sorted list. For these nodes, we again provide the popularity value of the nodes to the algorithm. Consider that with this strategy the algorithm also navigates greedy by degree.

The rationale behind the clustering strategy for node selection is that nodes with low clustering reduce the probability of a link to exist between two random neighbors of a node. This means that selecting a node with low clustering will provide nodes where the neighbors are not connected. The absence of a link between two neighbors of a node can be interpreted in the way that the neighbors are just too different. This would imply that selecting nodes with low clustering would provide nodes whose similarity between the neighbors is very small and this would allow navigation between clusters in the network. On the other hand, low clustering means that in this network region there is a *structural hole* as defined by Burt in [5]. The absence of connections between the nodes in these regions of the network will give even a higher importance to the existing connections resulting in a higher importance of the nodes in these regions.

3.3 Evaluating Navigational Efficiency

As emphasized, we conduct experiments with two distinct strategies for selecting the node members of the informed set having also different informed set sizes. With the popularity and clustering strategy (see Section 3.2), we examine how the exposure of the structure of the information space through the interface affects the efficiency of navigation. Furthermore, with the size of the informed set, we investigate how much structural information is needed for efficient navigation.

We conduct experiments on four different networks (see Section 4): (i) Wikipedia for schools (topological link network), (ii) Facebook (ego network), Twitter (ego network) and (iv) DBLP (co-authorship network). The four datasets can be seen as representatives of popular networks in information system on the web. For each network, we generate thousand navigational missions containing of one *start node* and one *target node* chosen randomly with at least one path between them. The goal for the algorithm is to reach the target nodes. We break up the search after 20 iterations on the small networks (Wikipedia for schools and Facebook) and



(a) Wikifs (b) Facebook (c) Twitter (d) DBLP

Figure 3: **Degree distributions on log scaled axes.** We see that all networks have a power law like degree distributions. This implies that degree greedy navigation will be very successful. For α-values cf. Table 2.

50 iterations on the big networks (Twitter and DBLP). We conduct experiments with degree as local popularity metric.

Note that for a set of size 0% of all nodes in the network, we navigate without any structural information. With this setting, the partially informed decentralized search reduces to a uninformed random walker (cf. Figure 2) which can serve as a baseline for our experiments as it corresponds to a third (random) strategy for selecting important nodes. For a set size of 100%, we navigate with all available information. This means that the partially informed search upgrades to a fully informed decentralized search.

As we are interested in examining the impact of the amount and kind of structural information provided to the partially informed decentralized search algorithm, we also need to evaluate the efficiency of the algorithm. To that end, we focus on two metrics: the *success rate* and the *stretch*. Success rate and stretch respectively measure the effectivity and efficiency of the search. We calculate the success rate as:

$$s = \frac{|W|}{|P|} \tag{3}$$

It is the fraction of the set of successful missions W and the set of all missions in the simulation P. The success rate measures the percentage of cases in which the algorithm was able to find the target node. Thus, the success rate measures the effectivity of the algorithm. To measure the efficiency of the algorithm, we consider the stretch defined as:

$$\tau = \frac{1}{|W|} \sum_{s,t \in W} \frac{h(s,t)}{l(s,t)}. \tag{4}$$

Technically, the stretch is calculated by dividing the length of the path produced by the algorithms $h(s,t)$ with the length of the shortest path $l(s,t)$ between the start and the target nodes and then averaging over all nodes.

4. DATASET DESCRIPTION

In this section, we give a thorough description of the studied datasets and their structural properties. We analyze four different networks (cf. Table 1) taken from the Stanford Large Network Dataset Collection[1]. The *Wikipedia for schools* network represents the topological hyperlink network derived from Wikipedia articles for teaching purposes referred to as Wikipedia for schools (Wikifs). The *Facebook* and *Twitter* datasets are ego-networks. Finally, the *DBLP* dataset represents a co-authorship network.

Table 1: **Datasets collection.** The table shows the network type and the number of nodes and edges. Two networks are directed and two undirected. For each network type there is a small and a big network regarding the nodes and the edges.

Name	Type	Nodes	Edges
Wikifs	directed	4,604	119,882
Facebook	undirected	4,039	88,234
Twitter	directed	81,306	1,768,149
DBLP	undirected	317,080	1,049,866

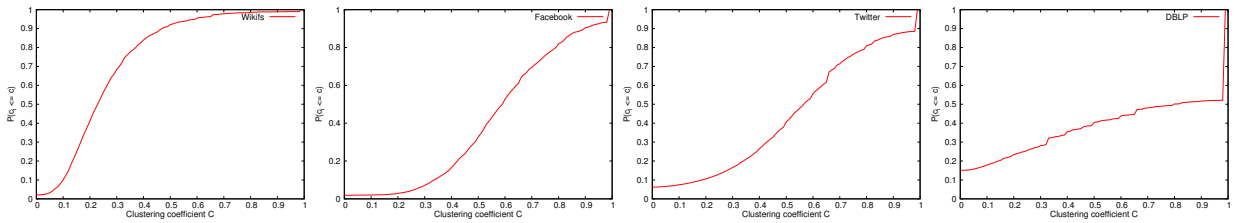
Navigability of networks. In Figure 3, we see the degree distributions of the different datasets. All networks exhibit power law like degree distributions at least for the tail. To get an initial idea of the navigability of these networks, we apply the method presented by Boguña et al. [4] who studied navigability of networks by looking at their clustering coefficients and power law exponents. In Table 2, we see that the values of the clustering coefficient of all networks are in the range defined in [4]. Additionally, we determine the power law exponent of the degree distributions with the methods presented in [6, 3]. We see that if we try to fit the power law distribution for the whole range of data points ($x_{min} = 1$), all networks are navigable according to Boguña et al. [4]. This is not the case, if we try to find the best power law fit and let the method estimate the best x_{min}. In this case, only the Facebook network is efficiently navigable.

Inequality of degree distributions. The Gini index is a metric that reviews the inequality in the degree distributions. A Gini index of zero means that the degree is equally distributed over the network, whereas a Gini index of one means that one node of the network possesses all links. In Table 3, we highlight the Gini index and the corresponding functions generating distributions with such inequality for the four datasets at hand. The corresponding generating functions support the results of the estimated first data point. We see that Wikipedia for schools, Facebook and DBLP possess Gini indices of 0.54. The inequality in the degree distribution is more explicit in the Twitter network. Inequality in the degree distribution is important for achieving good results with greedy navigation since it assures easy decision making.

[1] `http://snap.stanford.edu/data/index.html`

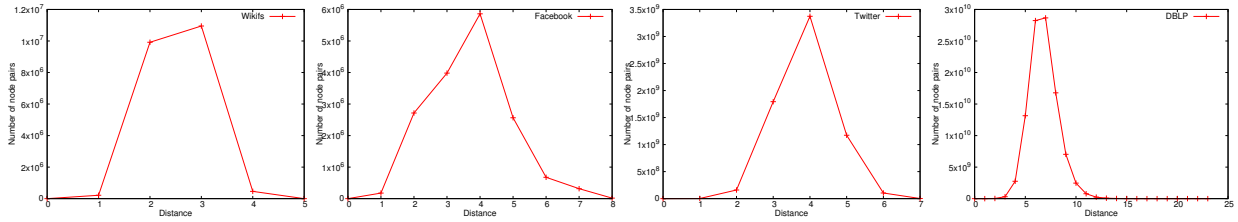
Table 2: **Small world classification of the datasets.** Depending on the point from where we try to fit the power law in the distribution (from the first data point or x_{min} estimated automatically), we see that either all of the networks are efficiently navigable (the clustering coefficient C and the power law exponent α are in the range defined by Boguña et al. [4]) or just the Facebook network. Table 3 suggests that we have higher trust in the results of the second row where the x_{min} is placed automatically.

Network	C	α, x_{min}	SW?	α, x_{min}	SW?
Wikifs	0.27	1.25, 1	✓	3.05, 142	✗
Facebook	0.61	1.26, 1	✓	2.51, 47	✓
Twitter	0.57	1.30, 1	✓	3.27, 188	✗
DBLP	0.63	1.48, 1	✓	3.26, 29	✗



(a) Wikifs (b) Facebook (c) Twitter (d) DBLP

Figure 4: **Clustering coefficient distribution.** Most nodes in Wikipedia for school have clustering around 0.2 meaning that the network has no clearly defined clusters. Facebook and Twitter exhibit similar distributions despite the different network size; there is a fraction of nodes with clustering near zero and a bigger fraction of nodes with very high clustering near one. All other nodes have clustering coefficient nearly uniformly distributed between zero and one. Very characteristic for the DBLP network is the high clustering coefficient; around half of the nodes have a clustering coefficient of around one.



(a) Wikifs (b) Facebook (c) Twitter (d) DBLP

Figure 5: **Shortest distance distributions.** Most of the node pairs in Wikipedia for schools have very short shortest paths; this makes this network very efficiently navigable. We see that the Facebook and Twitter networks have very similar distributions despite the different network size. Also, in these networks most of the node pairs have very short shortest paths. DBLP is the most difficult to navigate considering the fraction of node pairs with relatively long shortest paths.

Pareto principle. Since all networks possess power law like degree distributions, the Pareto principle suggests that we will need at least 20% of the nodes to achieve similar success rates and stretches for the networks with the popularity strategy and partially informed decentralized search as with a fully informed decentralized search. Additionally, we see that only one network is navigable according to the classification of Boguña et al.[4] (if we use higher x_{min} values), thus, we cannot necessarily expect the popularity strategy with smaller amounts of nodes to perform well in these networks. The results presented in Section 5 contradict this intuition. We believe that this is tightly related to the clustering coefficient distributions for the four networks.

Differences in clustering coefficient distributions. In Figure 4, we see that the networks possess very different profiles regarding the clustering coefficient distributions. We see that the Facebook and Twitter networks exhibit similar clustering coefficient distributions, despite the different network size. In these networks, most of the nodes have a clustering coefficient between 0.3 and 0.7. Nodes in DBLP exhibit very high clustering coefficients and most of the nodes in Wikipedia for schools have a clustering coefficient between 0.1 and 0.5. Thus, we also expect to see differences in the results produced by the clustering strategy for node selection.

Shortest path distributions. Beside the clustering coefficient and the degree distribution of a network, the shortest distance distribution is also important for the emergence of the small world property of a network [31] which significantly increases its navigability. The shortest distance distribution also provides insight into how difficult it generally is to navigate a network. In Figure 5, we can see the shortest distance distributions of studied networks. For the Wikipedia for schools network, we see that most of the node pairs have a shortest distance of three. The Facebook and Twitter network exhibit a bit longer shortest distance, whereas DBLP has the longest shortest distance distribution.

Table 3: **Gini Index.** The table shows the Gini index of the used networks and the corresponding distribution functions.

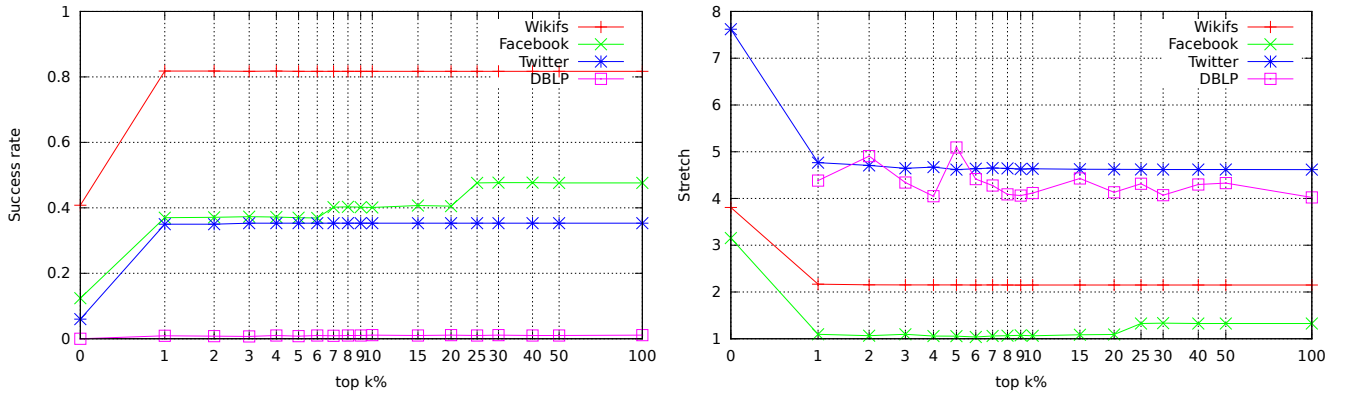
Network	Wikifs	Facebook	Twitter	DBLP
Gini Index	0.54	0.54	0.64	0.54
$f(x)$	x^2	x^2	x^3	x^2

5. RESULTS

In the following, we provide the results of our empirical evaluation. For both node selection strategies presented in Section 3.2, the decentralized search algorithm navigates greedy by degree. The amount of information needed for efficient navigation depends on the type of the structural information used and differs in the distinct networks.

Popularity strategy results. First, the popularity strategy tries to identify important nodes based on their popularity. Figure 6 shows the success rate and stretch for this strategy in all networks. In this case, the algorithm achieves with just 1% of the nodes similar efficiency results as with 100%. For Facebook, the partially informed decentralized search achieves slightly worse results than the fully informed search already with 2% for navigation by degree and the same or even a bit better results with 25% of the nodes. For this setting, the algorithm achieves similar performance as the fully informed decentralized search for Wikipedia for schools and Twitter also already with 1% of the nodes. We see that navigation in DBLP is very difficult in general. The best results in this network are realized with 2-3%.

Clustering strategy results. The clustering strategy tries to identify structurally important nodes based on their clustering coefficient. Figure 7 shows the success rate and stretch for greedy navigation



(a) Success rate

(b) Stretch

Figure 6: **Success rate** (s) **and stretch** (τ) **for popularity strategy for different amount of information.** *Left (a):* The success rate achieved for the popularity strategy and degree as popularity metric—the higher the better. To improve readability, we added one to all values and logarithmically scaled the x axis which shows the amount of information used. *Right (b):* The stretch achieved for the popularity strategy and degree as popularity metric—the lower the better. To improve readability, we added one to all values and scaled the axes logarithmically. We can see that we can achieve the success rate and stretch levels of fully informed search already with very small amount of information—about 1-2%. Strongly outperforming the fully informed search is not possible with this strategy.

by degree. We see that for Wikipedia for schools and Twitter the success rate initially falls with increasing amount of information, and then it jumps to the level of the fully informed search at 2% and 6% for Wikipedia for schools and Twitter, respectively. For Facebook, we observe very interesting success rate values since we are able to achieve considerably better results with less structural information. The success rate grows from 1% to 6% of the nodes to a value higher than the value achieved by the fully informed search (100%). After a drawback between 6% and 9% of the nodes, the success rate achieves even better results than for 6% with 15% of the nodes. Using more than the top 15% of the nodes worsens the success rate to the level of fully informed search. As before, we can see that navigation in DBLP is also very difficult with this strategy. The best results in this network are realized with 30% of the structural information.

Findings. Next, we summarize the results in the following two main findings answering the research questions tackled throughout this work as proposed in Section 1.

(i) What kind of structural information is needed for efficient navigation? Strongly outperforming the fully informed search with the popularity strategy is not possible. With increasing amount of information about the popularity, the success rate and the stretch improves continuously. With the clustering strategy, it is partly possible to substantially outperform the fully informed search. There is an initial drawback in success rate and stretch in all networks with the clustering strategy. After this initial drawback the success rate and the stretch increase until the levels of the fully informed search or even outperform the fully informed search.

Finding 1: Our results suggest that nodes with high popularity and low clustering are very important and can guide navigation very well and thus, should be exposed to the user through the interface.

(ii) How much structural information is needed for efficient navigation? With the popularity strategy, the levels of success rate and stretch produced by the fully informed decentralized search are achieved already with 1% of the popularity information. With the clustering strategy, the levels of success rate and stretch produced by the fully informed search are achieved with a bit more information than with the popularity strategy, depending on the network.

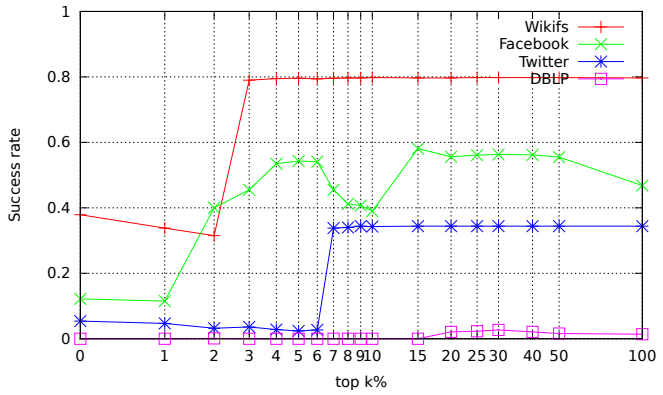
Finding 2: Our results suggest that with intelligent selection of nodes based on their structural properties, we can significantly reduce the amount of information that is needed to be presented to the user in navigational interfaces without reducing the efficiency of navigation.

6. DISCUSSION

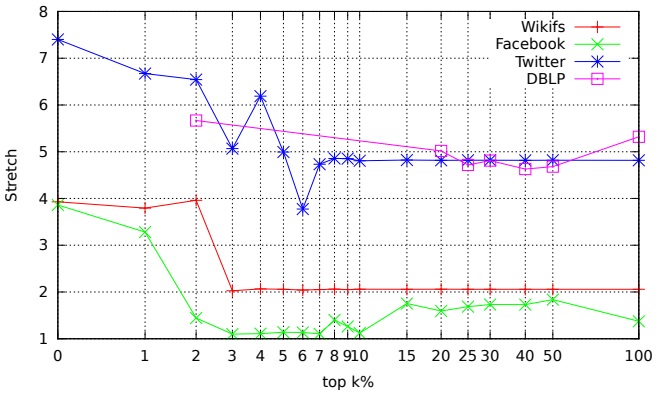
In Section 6.1, we start with a discussion and interpretation of our results (cf. Section 5) tailored around the research questions posed in Section 1. In Section 6.2, we discuss the implications followed by an elaboration of the advantages and limitations of our approach in Section 6.3.

6.1 Discussion and Interpretation of Results

Quality of Structural Information—Popularity vs. Clustering. Ranking the nodes by popularity and clustering is a good way to identify structurally important nodes. Furthermore, if the popularity information is combined with small amounts of clustering information which is a local metric, we can navigate even more efficiently. Nodes with high popularity and low clustering are very important and can guide navigation very well and should be exposed to the user through the interface. Knowing the important nodes on the local level regarding popularity and clustering can result in reducing the amount of nodes that need to be exposed to the user. This way we would be able to relax constraints of the screen size [12]. The initial drawback in the performance of the algorithm for this strategy can be explained by the degree distributions of the informed set of nodes. If the set is too small, there are not enough nodes with high popularity. Once the informed set has a sufficient amount of nodes for which the user has an intuition not only about the popularity but also about the clustering coefficient of the nodes, the user can navigate more confidently towards the target.



(a) Success rate



(b) Stretch

Figure 7: **Success rate** (s) **and stretch** (τ) **for clustering strategy for different amount of information.** *Left (a):* The success rate achieved for the clustering strategy and degree as popularity metric—the higher the better. To improve readability, we added one to all values and logarithmically scaled the x axis which shows the amount of information used. *Right (b):* The stretch achieved for the clustering strategy and degree as popularity metric—the lower the better. To improve readability, we added one to all values and scaled the axes logarithmically. We can see that for achieving the success rate and stretch levels of fully informed search (100%) we need slightly more information with the clustering strategy compared to the popularity strategy presented in Figure 6. Nonetheless, with this strategy, we are also able to outperform the fully informed search in some networks by only utilizing a low amount of clustering information.

Amount of Structural Information—Partially vs. Fully Informed Search. Surprisingly low amount of structural information is needed to achieve the same or even better results than with all information. This finding is really surprising if we consider the level of inequality in the degree distributions suggested by the Gini index and the exponent of the power law degree distribution (cf. Table 2 and Table 3). For the popularity strategy, outperforming the fully informed search is not possible, whereas for the clustering strategy we are able to top the results produced by the fully informed decentralized search.

6.2 Implications

Navigation in online networks is supported by smart user interfaces like tag clouds, breadcrumbs, subcategories and related categories. Normally, these navigational user interfaces make use of algorithmically preprocessed information about the content of the network. Our results have direct implications for these algorithms and for the ways data is presented to the user through the navigational interfaces.

Rethinking algorithms. Our findings suggest to reorganize the way we build hierarchies and to rethink algorithms creating hierarchies like [15, 12, 26, 7, 32]. In [12], the authors showed that the ability of hierarchies to guide navigation is significantly reduced through the restrictions introduced by the user interfaces. The main problem identified by the authors was that the top level of the hierarchies produced by the algorithms have too many subcategories—i.e., a too high *branching factor*. To tackle this problem, they adapted one of the best known algorithms for hierarchy induction proposed by Heymann and Garcia-Molina [15]. This algorithm creates a hierarchy by producing a similarity network. The hierarchy is then developed by ranking the nodes in the similarity network by popularity. Nodes are then placed in the hierarchy in a descending order of their popularity and their similarity to nodes that are already in the hierarchy. This way, nodes with high popularity are placed in the top of the hierarchy and nodes with low popularity at the bottom. Our results suggest an alternative ranking for the nodes of the similarity network. The results of the popularity strategy suggest that we should concentrate on the top 1% of the nodes in the network and try to produce a hierarchy with a well structured top. In contrast, the order of the bottom levels of the hierarchy is not really important, since we are able to achieve the same efficiency in navigation with only 1% of the nodes. This result also suggests that even if we break the semantics in the low levels of the hierarchy, we still will be able to navigate efficiently. Hints of how we should reorganize the top levels of the hierarchy are given by the clustering strategy we presented. We can re-rank the nodes of the similarity network considering not only their popularity, but also their clustering coefficient.

Our result could also be applied to the adapted version of the algorithm by Heymann and Garcia-Molina [15] proposed by Helic and Strohmaier [12] which generates a hierarchy in two stages. First, it produces hierarchies with a given branching factor. The largest hierarchy is called the main tree and all other hierarchies are then added to the main tree. After sorting the hierarchies by size, they are attached to the main tree in a way that preserves the branching factor of the hierarchy. Here, we could again try to re-rank the most popular nodes also by their clustering and put them in the main tree as suggested by the clustering strategy.

Presentation and information scent. Our results suggest that for efficient navigation, only a very small amount of local popularity and clustering information is necessary. Thus, we can derive that for efficient navigation, the user needs to have a good intuition only about the most important nodes in the network. Exposing the nodes with high structural importance through the user interface does not ensure that the user is going to select them. If the user has no sufficient knowledge and understanding of the most important nodes, the system has to deliver the explanation and in this way strengthen the *information scent* of the user for these specific nodes [24, 25]. By providing additional information about the important nodes regarding popularity and clustering, the information system would help the user to create an intuition about the presented choices. Assuring that a user has a high understanding about the topology of the information space—high exposure to the structurally most important nodes—would allow us to reduce the actual amount of nodes that are presented to the user though the interfaces. Without such information, random navigation performs well, which is con-

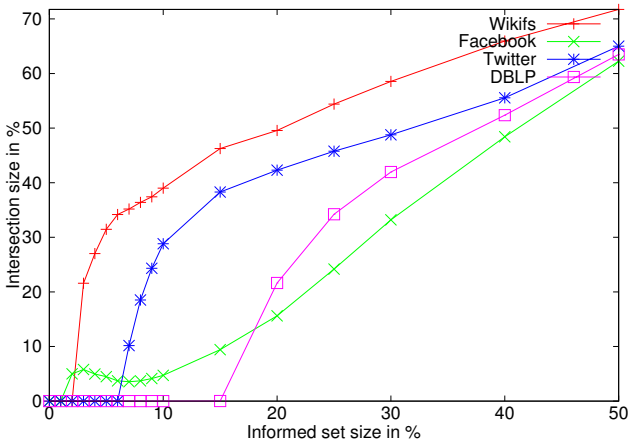


Figure 8: **Informed sets overlap.** The intersection size in percentage for the popularity and clustering strategy for the different informed sets sizes (except for 100%, $I = V$) of nodes used in the experiments in Section 3.3. For the datasets at hand, there is no big overlap in the informed sets selected by the two strategies. Additionally, with increasing set size the overlap does not necessarily increase as so the efficiency of navigation (cf. Figure 6 and Figure 7).

sistent with previous results for navigation by popularity in power law networks [2].

Helic et al. [14] studied the navigability of social tagging systems and showed that the tagging networks are power law networks. They showed that limiting the tag cloud size to practically feasible sizes (e.g., 5%,10%) does not affect the navigability. Our results suggest that we can reduce the tag cloud size even further to 1% of the nodes, according to the popularity strategy. In the same work, the authors also provided theoretical and empirical arguments against existing approaches of tag cloud construction. Possible improvements of these approaches can be achieved for instance with alternative rankings considering the clustering of the tagging network as the results of the clustering strategy presented.

6.3 Advantages and Limitations

In the following, we would like to address some limitations and advantages of our work.

Correlation between strategies. It has been shown that networks might exhibit a negative correlation between the degree and the clustering coefficient of nodes based on the formal definition of the clustering coefficient [28]. Due to this negative correlation, it is possible that there is big overlap in the informed sets created by the popularity and clustering strategy in this work. That is why it is important to quantify up to which extent the two strategies for important node selection differ in the experiments conducted in Section 3.3. In Figure 8, we illustrate the size of the intersection of the popularity and clustering strategies for all datasets for different sizes of the informed sets of nodes. Overall, we can see that the overlap of nodes between the two strategies is considerably low for smaller set sizes. Not surprisingly, with increasing set size, the overlap is generally rising as the chance of overlapping node selection increases. However, as it can be seen in Figure 6 and Figure 7, an increasing overlap does not reflect an increase in the performance of the partially informed decentralized search. By and large, these observations support the importance of the findings from Section 5 and give a confirmation that both strategies select mostly different nodes and structurally important nodes that could support navigation.

However, there might exist some few nodes that are highly beneficial to be included in an informed set for efficient navigation. Both strategies might select them early on and as soon as they include these nodes, efficiency increases drastically. Thus, in future work, we plan on further investigating the overlaps between both strategies which might also help us to find even better (potentially smaller) informed sets that can guide navigation well.

Alternative strategies for node selection. With our experiments, we have concentrated on the degree and clustering coefficient as metrics for measuring the structural importance of nodes. Above, we have discussed the potential correlation between both strategies but have also shown that the overlap is low for small informed sets. Nonetheless, other strategies might be amendable. For example, previous work [28] has suggested an alternative way to calculate the clustering coefficient by removing the degree bias (cf. Equation 1). By utilizing this method, we might be able to further investigate the differences of both strategies for finding important nodes for navigation. Also, we could simply try to implement strategies for node selection that produce mostly distinct sets of nodes. By doing so, we might be able to further improve our approach potentially leading to even better results in terms of success rate and stretch. Nodes of the distinct sets selected by the strategies can then be exposed to the user through the navigational interfaces. Also, there exist other thinkable metrics (e.g., k-core and link irregularity) describing the structure of a network that can be applied in straightforward fashion [8, 9]. We leave these investigations open for future work.

Alternative user models. In our experiments we utilized a greedy neighbor selection if at least one of the candidate nodes is in the informed set of nodes otherwise we selected one at random. This models a user who always follows her intuition if it has one. Although this is a valid user model it is a very simple one. In future work we plan to experiment with alternative neighbor selection mechanisms that model a user who is greedy or stochastic to different extents in following her intuition [13]. Additionally, it is also thinkable to use different informed sets at different stages of the search e.g., the informed set created with the popularity strategy can be used in the beginning of the search where the user is interested in exploring the information space, whereas the informed set created with the clustering strategy can be applied in stages of the search where bridging a gap between two clusters is needed.

Global information. One limitation of the decentralized search is the amount of global information used for navigation. The models presented by Watts et al. [30], Kleinberg [18, 19] and Boguña et al. [4] make use of the global position of the target node. One could argue that partially informed decentralized search is using to much global information in the sense that it uses the information about the distribution of the degree and clustering coefficient. A way to tackle the problem would be to make a random sample of n% (i.e., 30%) of the nodes and apply the popularity and the clustering strategies only at these n% of the nodes in the network.

7. CONCLUSION

Navigational interfaces are only useful for augmenting navigation to the extent to which they are able to expose the underlying structure of the information space. In this paper, we have been interested in studying (i) which and (ii) how much structural information is necessary for properly exposing the hidden structure of the information space. To that end, we have utilized an adapted version—i.e., partially informed—of the message passing decentralized search algorithm. This adaption allows to model a user that is limited in

her exposure to the structure of the information space having only limited knowledge about the topology of the information space. In detail, we have focused on two strategies for selecting important nodes based on their (i) popularity and (ii) clustering coefficient.

With simulations on four distinct datasets, we have observed that a surprisingly low amount of structural information is needed by the partially informed version of decentralized search in order to achieve the same or even better performance than the fully informed decentralized search. Besides the popularity, for choosing structurally important nodes, also the clustering coefficient has turned out to be a good indicator for this task. The clustering strategy would expose nodes of high structural importance to the user which can be used to reduce the amount of information offered to the user and relax constraints posed by the limited size of the screen. Our results have implications on the algorithms used for the structuring of the information space. These algorithms should take into account the supportive properties of the clustering coefficient for navigation.

In future work, we would like to propose and evaluate another version of decentralized search that models exploitation on the local level. In this version, we plan on combining centrality metrics as proxies for popularity and clustering information as a proxy of homophily. With this extended version, we would like to study how the clustering coefficient can be used to jump from one network region to another or to stay in the same cluster and explore it.

Acknowledgments. This work was partially funded by the DFG in the research projects "PoSTs II" and "dalraSearchNet" (SU 647/13-2) and by the FWF Austrian Science Fund research project "Navigability of Decentralized Information Networks" (P24866).

8. REFERENCES

[1] L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27, 2005.

[2] L. Adamic, R. Lukose, A. Puniyani, and B. Huberman. Search in power-law networks. *Physical Review E*, 64, 2001.

[3] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9, 2014.

[4] M. Boguna, D. Krioukov, and K. C. Claffy. Navigability of complex networks. *Nat Phys*, 5, 2009.

[5] R. S. Burt. *Structural holes: The social structure of competition*. Harvard University Press, 2009.

[6] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM Review*, 51, 2009.

[7] I. Dhillon, J. Fan, and Y. Guan. *Efficient Clustering of Very Large Document Collections*. Kluwer Academic Publishers, 2001.

[8] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. K-core organization of complex networks. *Physical Review Letters*, 96, 2006.

[9] E. Estrada. Quantifying network heterogeneity. *Physical Review E*, 82, 2010.

[10] M. A. Hearst. *Search User Interfaces*. Cambridge University Press, 2009.

[11] D. Helic, C. Körner, M. Granitzer, M. Strohmaier, and C. Trattner. Navigational efficiency of broad vs. narrow folksonomies. In *Proceedings of the Conference on Hypertext and Social Media*. ACM, 2012.

[12] D. Helic and M. Strohmaier. Building directories for social tagging systems. In *Proceedings of the International Conference on Information and Knowledge Management*. ACM, 2011.

[13] D. Helic, M. Strohmaier, M. Granitzer, and R. Scherer. Models of human navigation in information networks based on decentralized search. In *Proceedings of the Conference on Hypertext and Social Media*. ACM, 2013.

[14] D. Helic, C. Trattner, M. Strohmaier, and K. Andrews. On the navigability of social tagging systems. In *Proceedings of the International Conference on Social Computing*. IEEE, 2010.

[15] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, 2006.

[16] W. Ke and J. Mostafa. Strong ties vs. weak ties: Studying the clustering paradox for decentralized search, 2009.

[17] W. Ke and J. Mostafa. Scalability of findability: Effective and efficient ir operations in large information networks. In *Proceedings of the International Conference on Research and Development in Information Retrieval*. ACM, 2010.

[18] J. Kleinberg. Navigation in a small world. *Nature*, 406, 2000.

[19] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the Symposium on Theory of Computing*. ACM, 2000.

[20] J. Kleinberg. Small-world phenomena and the dynamics of information. *Advances in neural information processing systems*, 1, 2002.

[21] J. Kleinberg. Complex networks and decentralized search algorithms. In *Proceedings of the International Congress of Mathematicians: invited lectures*, 2006.

[22] S. Milgram. The small world problem. *Psychology Today*, 61, 1967.

[23] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[24] P. Pirolli. Computational models of information scent-following in a very large browsable text collection. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 1997.

[25] P. Pirolli and S. Card. Information foraging. *Psychological Review*, 106, 1999.

[26] A. Plangprasopchok, K. Lerman, and L. Getoor. Growing a tree in the forest: Constructing folksonomies by integrating structured metadata. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. ACM, 2010.

[27] C. Seifert, B. Kump, W. Kienreich, G. Granitzer, and M. Granitzer. On the beauty and usability of tag clouds. In *Proceedings of International Conference on Information Visualisation*. IEEE, 2008.

[28] S. N. Soffer and A. Vazquez. Network clustering coefficient without degree-correlation biases. *Physical Review E*, 71, 2005.

[29] C. Trattner, P. Singer, D. Helic, and M. Strohmaier. Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks. In *Proceedings of the International Conference on Knowledge Management and Knowledge Technologies*. ACM, 2012.

[30] D. Watts, P. Dodds, and M. Newman. Identity and search in social networks. *Science*, 296, 2002.

[31] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393, 1998.

[32] S. Zhong. Efficient online spherical k-means clustering. In *Proceedings of the International Joint Conference on Neural Networks*. IEEE, 2005.