

Information access representations and social capital in networks*

Ashkan Bashardoust University of Utah Salt Lake City, UT, USA ashkanb@cs.utah.edu	Hannah C. Beilinson Haverford College Haverford, PA, USA hcbeilinson@gmail.com	Sorelle A. Friedler Haverford College Haverford, PA, USA sorelle@cs.haverford.edu
Jiajie Ma Duke University Durham, NC, USA jason.ma@duke.edu	Jade Rousseau Haverford College Haverford, PA, USA jade.rousseau.paris@gmail.com	
Carlos E. Scheidegger Posit PBC USA carlos.scheidegger@posit.co	Blair D. Sullivan University of Utah Salt Lake City, UT, USA sullivan@cs.utah.edu	
Nasanbayar Ulzii-Orshikh University of Michigan Ann Arbor, MI, USA nulziior@umich.edu	Suresh Venkatasubramanian Brown University Providence, RI, USA suresh@brown.edu	

October 17, 2023

Abstract

Social network position confers power and social capital. In the setting of online social networks that have massive reach, creating mathematical representations of social capital is an important step towards understanding how network position can differentially confer advantage to different groups and how network position can itself be a source of advantage. In this paper, we use well established models for information flow on networks as a base to propose a formal descriptor of the network position of a node as represented by its information access. Combining these descriptors allows a full representation of social capital across the network. Using real-world networks, we demonstrate that this representation allows the identification of differences between groups based on network specific measures of inequality of access.

1 Introduction

It has been known for decades that belonging and positionality in a social network confer power to individuals in the form of *social capital* [16, 17, 18, 24, 25, 36, 40, 43, 61]. More recently, information access has appeared as one of, if not *the*, key component of social capital. On the one hand, information access is structural in nature such that what information one has access to is determined by who one knows, by one’s place in social networks. On the other hand, what information one has access to (in a broad sense of the term) is so determining for navigating the social world that position in a network is itself a first-order “feature” that can lead to discrimination in and of itself, separately from (while often compounded with) individual demographic features [14]. Both the structural nature of information access and its relation to overall individual social capital is exacerbated in *online* social networks, where curation processes – recommendations most notably

*This research was funded in part by the NSF under grants IIS-1955321 and IIS-1956286.

– encourage the formation of self-preferential links between people (nodes) and thereby reinforce existing information advantage and help create classes of people based on their relative access to information. Node groups based on information access might therefore be more salient to understanding network privilege dynamics than traditional groups based on node-level features like demographics (a.k.a sensitive attributes). It is this observation that motivates this work. Our goal is to formalize a representation of social capital, based on information access, for a network such that individuals’ social position and associated group structure can be studied.

1.1 Contributions

In this paper, we:

- connect the theory of social capital and information access to an introduced mathematical representation;
- formalize the notion of an *information access representation* of a network;
- show how a full representation can be effectively computed for a variety of network sizes;
- and validate the representation by showing that clustering nodes based on information access captures external measures of real-world interest through experiments on real-world networks.

We also present a formal interpretation of information access representations and distinguish it theoretically and empirically from existing methods of identifying structure in networks, illustrating the novelty and value of the information access lens.

2 Related Work

Inspired by the literature on social position initiated by [37] and framed in the context of online social networks by [14], there has been a recent development of computational questions around fairness in access on social networks [4, 9, 10, 31, 45, 62, 73]. The starting point for all of this is the idea of *information access as a resource*. In [31], this is formalized in terms of the question “what is the probability that vertex i in a graph gets information from source j ,” and the paper focuses on explicit interventions (in the form of augmenting the seed set) to ensure that the minimum such probability is maximized. Subsequent research [70, 73] has focused on the problem of allocation with respect to information access – attempting to make sure that different demographic groups (represented as disjoint subsets of the graph) receive similar information access. Recently, this work has expanded to include structural notions of such fairness of access [44, 54], still with a focus on demographic groups. Information access itself draws on the mechanisms of influence maximization, which was studied initially by [28] and formalized by [50], leading to an extensive literature on the subject [52]. Related literature that is not directly connected to this work includes an examination of network formation biases (through natural mechanisms and through automated recommendations in online social networks) might exacerbate inequity (in the form of reduced social capital) and reduced opportunities for those in the out-groups [23, 49, 51, 56, 58, 70, 78, 80]. Another line of work [8, 45, 71] consider adding edges instead of seeds to improve fairness in networks. In particular, Bashardoust et al. measures the structural advantage of a node based on the access signatures [8], referencing the originally preprinted version of this paper [11].

The idea of using social processes of group formation in a method for finding *clusters* in a graph, while distinct from our information-access-based framework, is generally referred to as *community discovery*. A slightly different perspective often referred to as *role-based* discovery also seeks to recognize that groups may not be identified based on proximity, but based on similar *roles* that entities play in the network. For example, Henderson et al. propose role extraction models in large graphs [38, 39]. These are not mutually exclusive notions: in practice one way to “interpolate” between the two approaches is to decide the size of neighborhood that is relevant when determining if two nodes have similar roles or are in the same community. See [64] for a recent survey of this literature. Another line of work recognizes that graphs manifest a variety of structures at the *mesoscale*[32] – when we are not merely looking at either individual node neighborhoods or

aggregate properties of the entire graph. Thus, a fruitful strategy is to focus on specific structures of interest and detect them directly. One such pattern that is particularly relevant for this paper is *core-periphery* structures [13, 26, 63] described by a central well-connected core that links with a number of disconnected pieces. We explore connections to these methods further in Section 6.

Moving further afield, graph clustering itself has long been a focus of intense study (see [2] for a survey). While a detailed review of the different strands of graph clustering is beyond the scope of this paper, one can categorize graph clustering algorithms as those based on finding dense submotifs in a graph, those based on spectral analysis [34, 47, 68, 77] (which in turn generalizes connectivity-based clusterings) and those based on the more general framework of unusual local density or modularity [15].

Recent research has delved into the examination of the information access challenge within the realm of theoretical graph neural networks (GNNs) [5, 6, 35, 48, 72]. Within this body of work, a number of studies propose the utilization of random-walk spectral metrics as a means to effectively characterize the dynamics of information propagation [6, 35, 72, 75]. Additionally, Dong et al. address the problem of explaining and mitigating the structural bias of input networks in GNNs [29, 30].

3 Information Access: Social Theory

Social capital refers to the value that is gained from ‘being social’; that is, the idea that being a part of social structures is a source of benefit, utility, power; and, further, one’s place within these social structures is itself a determinant of power in part because it determines access to various resources [46]. Most of us likely already have an intuitive idea of the significance of belonging and positionality when it comes to access to information—e.g., who you are friends with determines whether you will be aware that an event will occur. This access to information is a resource determined by our relations with others [42].

Position in a network is thus a first-order ‘feature’. As such, it can in and of itself lead to discrimination though in reality, it is intertwined with other first order features such as individual demographics [14]. Such an emphasis on position in a network is characteristic of the *network approach* to social capital, which developed following Granovetter’s 1973 seminal paper [37]. This approach tends to focus on *structural social capital*: looking at the properties of the social system and of the network of relations as a whole, and approaching social capital based on what relationships (edges) reveal about a chosen proxy of social capital [20, 21]. We take this network approach to social capital, and acknowledge in doing so that we do not examine aspects of social capital that may not be visible through a focus on and examination of network structure. Information is also a locus of power, and various studies have linked deep-rooted inequalities to disparate levels of access to information [55], such as racial income differences in the US [42]. Information exists and is accessed *within* networks and one’s relationship to information is thus a crucial element of one’s social capital; a relationship which network analysis is uniquely positioned to investigate.

Furthermore, information is not merely spread through a social structure, it itself participates in the creation of that structure [7]. In online social networks, not only is information (in the form of content) essentially what these platforms are about, but the users themselves are, in an important way, defined in terms of their relationship to information: which content a user is exposed to, which information they consume, how they consume it, and so on itself becomes information as *data* [57]. We thus argue that information access can be conceptualized and analyzed not only as a resource an individual has or hasn’t access to, but as a defining feature of an individual. Our claim is consistent with a cultural structuralist conceptualization of social capital, which views social capital as not merely something an individual *has*, but as *a part of* an individual [22, 69].

We therefore propose to represent the social capital of a node i in a network in terms of its *information access signature*: a vector containing the probabilities of node i accessing information from each other node in the network j which spreads it. Introduced formally in Section 4.1, information access signatures for all nodes can be combined into a matrix representation of the information access of the full network. Because the matrix is symmetrical along the diagonal in an undirected network, the vector may also be thought of in terms of ‘influence’ (or ‘contagion’, ‘diffusion’): that is, as containing the probabilities of node i spreading an information to each other node j . The information access signature thus *structurally* defines a node in terms of the map of its access to information. For two nodes to have a similar information signature thus means that they have a similar pattern of connection, which in turn means that they are *structurally*

equivalent. The concept of structural equivalence was developed in sociology to explain the mechanisms of diffusion, with the goals of understanding how information comes to be spread or how a practice comes to be adopted, and is a characteristic of nodes that occupy similar positions and roles in a network [1]. Structurally equivalent nodes do not need to have linkages to each other, nor do they need to be connected to the same nodes for them to share a similar structural position; what matters is the *pattern* of connections [53]. For example, one might draw strong parallels between the conditions of people living in different cities, though they inhabit different networks, because of similarities in their *structural embedding*. We thus claim that the information access representation of the network is a mathematical representation of social capital that allows examination of structurally equivalent individuals or groups in a network, e.g. via clustering. We contend that since information access is both a critical first-order feature and a feature within which other features, such as demographics, are embedded, this representation offers a novel and valuable way of looking at communities and inequalities, problems and solutions.

4 Information Access: Mathematical Representation, Clustering, and Analyses

Let $G = (V, E)$ be a network with sets of nodes V and edges E , where $|V| = n$. Consider any *information flow model* that describes how information might transmit from one node to its neighbors, such as the independent cascade model, the linear threshold model, or any of the infection flow models from epidemiology. All these models are stochastic and assume some initial *seed* set of nodes that possess the information to be spread. For any given seed set S there is then a fixed probability $p_{v,S}$ that node $v \in V$ possesses the information once the process terminates. This motivates our idea of an information access signature: a way to encode the “view” from a node v of the access it has to information sent from other nodes in the graph.

4.1 Information access signatures, representation, and clustering

Definition 1 (information access signature). *The information access signature $s_\alpha^G: V \rightarrow \mathbb{R}^n$ for $v_j \in V$ is $s_\alpha^G(v_j) = (p_{1j}, \dots, p_{ij}, \dots, p_{nj})$ where p_{ij} is the probability that node $v_j \in V$ receives information seeded at node $v_i \in V$ under some information flow model with parameter vector α .*

Intuitively, this signature characterizes a node’s information access based on how likely they are to receive information from everyone else in the network; people who are likely to receive information from the same part of the network will have similar signatures. Given our understanding of information access as a measure of network privilege, this means two nodes with similar access signatures should have comparable information privilege.

We can use these signatures to define an information access representation of the network.

Definition 2 (information access representation). *The information access representation for $v_j \in V$ is*

$$R_\alpha^G = \{ s_\alpha^G(v_j) \mid v_j \in V \}$$

where α represents the parameters of the information flow model as before.

The information access representation R_α^G represents each vertex of the graph as a point in an n -dimensional space. Two points are close in this space if they have similar views of information flow to/from other nodes, i.e. similar information access privilege or social capital. This thus allows us to group together nodes based on this proximity, i.e. to compute a graph clustering.

Formally, we will endow this space with an ℓ_2 norm, which then allows us to use standard clustering techniques. Throughout this work, we’ll use k -means, though other standard clustering techniques can be easily applied to the representation. We will call the clustering resulting from applying k -means clustering to the information access representation for a given α and k desired clusters the *information access clustering* of a network.

While these representations are well-defined for any model of information flow, in order to describe examples, run experiments, and discuss computational complexity, we need to fix a mechanism for information spread. As such, for the remainder of the paper, we assume that information is flowing according to the

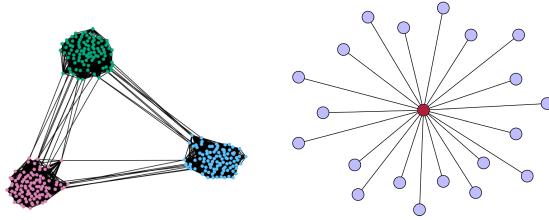


Figure 1: A graph of 3 communities created with the Stochastic Block Model (left) and a star graph (right) colored according to their resulting information access clustering (for $k = 3$ and $k = 2$, respectively).

extensively studied independent cascade model [50] (also known as the SIR model in epidemiology). In this model, a node exists in one of three states: ready to receive, ready to transmit, and dormant.¹ Initially, some subset of *seed* nodes possess a bit of information and are ready to transmit, and all other nodes are ready to receive. At each time step, a node that is ready to transmit iterates over all of its neighbors which are ready to receive, and sends the information with probability α (the decision is made independently for each adjacent edge). All such transmissions are imagined to happen simultaneously, after which the transmitting node goes dormant. Since the independent cascade model can be characterized by a single parameter α , in what follows we will merely use $s_\alpha^G(v)$ to denote the information access signature.

Unfortunately, as we show in Section 4.4, computing the information access signature under independent cascade is #P-hard. In order to get around this intractability for experiments, we estimate the probabilities using Monte Carlo simulations. Specifically, to estimate p_{ij} , we run S simulations of independent cascade where i is the only initial seed node and report the fraction of trials in which v successfully received the information; each simulation takes time $O(m)$ since an edge can only transmit information at most once in a given cascade. We observe that in order to compute the full representation, we need to do this with each of the n nodes in G as the initial seed, resulting in a time complexity of $O(Smn)$ and requiring space $O(n^2)$.

4.2 Examples

To demonstrate the goals and utility of the introduced representation and clustering we discuss both synthetic and real-world motivating examples.

Synthetic Graphs with Meso-scale Structure Two important identified mesoscale structures are block communities and core-periphery structures [63]. Since our proposed clustering is meant to capture structural information access patterns, we expect that in the first case, communities should correspond to clusters (members receive similar information with high access inside their own community and low access from those outside). In the second setting, the information access view should reflect the core-periphery structure, with central nodes having high access to information across the entire network and peripheral nodes all having similarly low access to information from anyone outside the core.

We now argue that on simple representatives from each of these classes, our approach does recover the desired clusterings (each community as a cluster, and a core versus periphery clustering, respectively).

We first consider a synthetic graph consisting of three strong communities of size 100, generated using the stochastic block model [41]. Each edge between members of the same community is present independently with probability 0.3, and edges between pairs of nodes in differing communities are present with probability 0.05. The information access representation matrix for this graph has a block structure with access probabilities averaging 0.88 between any two nodes in the same community, and average access probabilities between 0.62 and 0.67 between communities². Thus, the signatures for nodes in each community are highly similar (mostly differing at the entries corresponding to the two nodes being compared), but also easily distinguishable from signatures from any other cluster. We computed an information access clustering (sing the elbow

¹These states are called susceptible, infectious and recovered in the SIR formulation.

²These probabilities were estimated using 10,000 Monte Carlo simulations; the variance across individual nodes for the access between communities i and j was less than .001 in all cases ($1 \leq i \leq j \leq 3$)

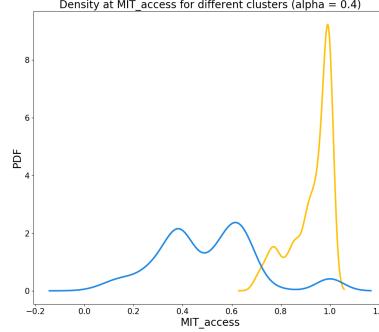


Figure 2: The probability density function showing the probability of access to someone working at MIT through the DBLP co-authorship network for two information access clusters. The average access probability for the blue cluster is 0.50 while the average access probability for the yellow cluster is 0.93.

method to determine the appropriate number of clusters $k = 3$), and observe that it does correctly identifies each community as a distinct cluster, as seen in Figure 1.

For core-periphery structure, we use a star graph which consists of a single central node connected to many peripheral nodes, each of which has no other neighbors. Any distinct pair of nodes in the peripheral set has probability α^2 of information transmission between them while any pair involving the central node has probability α . Thus, peripheral node signatures take the form $(1, \alpha^2, \dots, \alpha^2, \alpha)$ while the central node's signature is $(\alpha, \alpha, \dots, \alpha, 1)$. Again, if we select the number of clusters using elbow/silhouette, we find that the central node is in one cluster and all peripheral nodes are together in a second.

These examples demonstrate that the information access clustering we introduce is expressive enough to capture both community-based and core-periphery mesoscale structures.

Real-world example: DBLP co-authorship network We begin with an initial motivating experiment on the DBLP co-authorship network of scholars in computing-related disciplines (described in more detail in Section 5.1). What might information access-based social capital or privilege look like in a co-authorship network and how could it be measured? Given that the co-authorship network is an important form of social network within academia, we posit that one form of privilege in a co-authorship network is access to people in prestigious institutions. This access could provide field-critical information, i.e. the ability to both receive information *from* and share information *with* people at prestigious institutions, which might then impact faculty hiring (see, e.g., [23]). In an undirected network like the DBLP co-authorship network we study, these forms of access can both be modeled as the information access probability between people.

To examine whether the information access clustering groups scholars based on this probability of access to a prestigious institution, we take the case of MIT and examine the probability that an individual in the network is able to reach (or be reached by) someone working there. Using the same experimental setup described later in this paper and $k = 2$ clusters, we created the information access clustering and determined the probability of access to MIT for each individual. We find that one cluster is clearly the privileged cluster; its nodes on average have higher citation count, better job institution rank, and better Ph.D. institution rank. The average access probability across the privileged cluster is about two times that of the non-privileged cluster (see Figure 2), and the access to individuals from MIT is also statistically significantly ($p < 10^{-7}$) higher.³ This, coupled with our theoretical considerations, suggests that information access clustering may indeed allow investigation of information privilege or social capital within a network. With this motivation in mind, we next further consider the theoretical foundation of the representation and later consider a more comprehensive set of experiments (see Sections 5-7).

4.3 Spectral analysis

We now provide some formal insights into structures that information access clustering seeks to find, by looking more closely at the structure of the representation R_α^G and how it relates to structures in the

³These results hold across $\alpha \in [0.1, 0.7]$.

underlying graph G . For clarity, we drop the superscript G when the context is clear.

An alternate (and well-known) way to think about the independent cascade diffusion is as follows. Fix a graph G and define a distribution $\mathcal{D}_\alpha(G)$ of graphs as follows: delete each edge of $E(G)$ independently with probability $1 - \alpha$. Consider any graph H drawn from $\mathcal{D}_\alpha(G)$. Let $CC(H)$ denote the binary $n \times n$ matrix indexed by vertices of G such that $CC(H)_{ij}$ is 1 if v_i and v_j are in the same connected component of H . It is then straightforward to show that $R_\alpha = \mathbb{E}_{H \sim \mathcal{D}_\alpha(G)}[CC(H)]$. In other words, we can interpret p_{ij} as the probability that v_i and v_j are in the same connected component of a graph drawn from $\mathcal{D}_\alpha(G)$.

This interpretation also yields a connection to the spectral decomposition of G . Let D denote the diagonal matrix where $D_{ii} = \deg(v_i)$ and as usual let A denote the adjacency matrix of G . Then the (combinatorial) Laplacian of G is $L = D - A$. If G has ℓ components, then the nullspace of L is multidimensional, with one basis being the set of vectors v_1, \dots, v_ℓ where each v_i is a binary vector representing the characteristic vector of the i^{th} connected component. It then follows that $CC(G) = \sum_{i=1}^{\ell} v_i v_i^\top = V_G V_G^\top$ where V_G is the basis of the null space in which each of the characteristic vectors of the connected components is a column. Summarizing, $R_\alpha = \mathbb{E}_{H \sim \mathcal{D}_\alpha(G)}[V_H V_H^\top]$. This immediately indicates how our approach differs from (say) spectral clustering, which instead computes the eigendecomposition $L = U \Lambda U^\top$, fixes a parameter $d \leq n$ and then forms the representation $S = U_d$ where U_d is an $n \times d$ matrix containing the eigenvectors corresponding to the d smallest eigenvectors of L (and in particular *ignores* the nullspace).

4.4 Computation: Hardness & Simulation

Unfortunately, while the information access representation yields insight into the structure of the graph, it is intractable to compute exactly under the independent cascade model of information flow.

Theorem 3. *For an undirected graph G and a fixed probability of transmission α , the computation of each entry p_{ij} in the information access representation R_α^G is $\#P$ -complete when information is propagated via independent cascade.*

We recall that the entry p_{ij} in the signature is defined to be the probability that node v_j receives information seeded at node v_i . Shapiro et al proved that computing this quantity is $\#P$ -complete [67] under the SIR model with uniform transmission probability – which, as we have previously observed, is directly equivalent to independent cascade. Their proof is a reduction from the $\#P$ -complete *two-terminal network reliability problem* [74]: given a network G of n nodes and m edges where each edge of G is assigned a fixed probability $1 - \alpha$ that it disappears from graph (i.e., each edge survives with probability α), determine the probability p'_{ij} that in the surviving graph two particular nodes v_i, v_j are connected (have a path between them).

5 Experiments Part I: Social Capital

In order to provide validation that information access representations encode information about individuals' social position in a network as it relates to social capital, we examine the information access clustering produced on these representations for a variety of real-world networks. We will consider two main experimental questions:

1. Does information access clustering allow identification of clusters of individuals in the network that are structurally similar in terms of information access and real-world social capital?
2. How does the information access clustering compare to existing network clustering methods?

We begin with the first question in this section and return to the second in Section 6. Code to reproduce all experiments is available at: <https://github.com/algofairness/info-access-clusters/releases/tag/paper.1.0>

5.1 Network datasets

We choose real-world network datasets that share two fundamental characteristics. First, we have chosen networks based on domains where information tangibly flows across the network, and where access to that

information is clearly a form of privilege within that network. Second, all networks have an external node attribute that can be used to quantify information access or social capital within the network, which we will refer to as its *external information access measure*. By external, we mean that this attribute should not be directly encoded in the network structure itself; importance measurements such as node degree do not meet this criterion. Our goal with this criterion is to help answer our first experimental question and determine whether information access clustering, using solely the information access representation, clusters nodes together so that cluster composition is different with respect to the external information access measure. A clustering that separates nodes that are similar in terms of this external characteristic could allow identification of clusters that are advantaged or disadvantaged based on their information access without direct access to a measurement of that advantage, and in general could allow the unsupervised exploration of such external measures given only the information access representation. This would indicate that information about social capital is encoded in the information access representation. The networks studied range broadly in size - from 438 nodes in the Co-sponsorship dataset to 391,642 nodes in the Google scholar dataset - and experiments are run on the largest (strongly) connected component for each graph. Further considerations are taken into account for dealing with large graphs (see Section 7).

Dataset	#Nodes in Largest Connected Comp.	#Edges in Largest Connected Comp.	External Information Access Measures
DBLP	2,123	7,133	citation count, job rank, PhD rank
Twitch	7,126	35,324	partner, views
Co-sponsorship	438	28,194	legislative effectiveness score
Google Scholar	391,642	104,647,630	citation count, h-index

Table 1: Dataset information.

DBLP co-authorship In the DBLP network, nodes are scholars from <https://dblp.uni-trier.de/> and are connected by an edge if the scholars have co-authored a paper. This network contains only scholars who: received their PhD from a university on the Computer Research Association’s authoritative Forsythe List of Ph.D.-granting departments in computing-related disciplines in the United States and Canada; had their first assistant professorship at one of these universities; worked at one of these universities in the 2011-2012 academic year; and were hired between 1970 and 2011. This list of scholars, along with metadata about each of them, was compiled by [79]. Co-authors of these scholars were scraped from DBLP in October 2020. A small number of scholars were excluded either because their hire date was not known or because their DBLP id was inconsistent with the one recorded by [79]. The resulting network contains 2,356 nodes and 7,145 edges, with 2,123 nodes and 7,133 edges in the largest connected component.

The external information access measures used for this DBLP network are the number of total citations summed over all papers recorded by Google Scholar for a scholar (node) in the network⁴ and the ranks of the Ph.D. and first job institutions of the scholars (as recorded by [79], ordered so that 1 is the top rank).

Twitch The Twitch dataset is a social network from Twitch, a video live streaming platform for gamers, where the 7,126 nodes represent Twitch users who stream in English and 35,324 undirected edges are mutual friendships between them. The dataset was collected by [66] in May of 2018. Each node in the network is either a “partner” or not: a Twitch user becomes qualified to be a “partner” by accumulating streaming hours and maintaining more than 75 concurrent viewers on Twitch or having a sizeable audience on other platforms such as YouTube or Twitter. Thus, the external information access measures we use are the number of views for each user and their “partner” status.

Congressional co-sponsorship The Co-sponsorship data is a directed network indicating bill co-sponsorship for the 114th Congressional sitting in the United States. This dataset was first created by GovTrack and made available by [33] in July of 2017 after its removal from the original source. Based on this data, we

⁴Citation counts were scraped from Google Scholar in Spring 2020.

created a network in which each node is either a sponsor or original co-sponsor of a bill, and there exists a directed edge from one node to another if the latter originally co-sponsored at least one bill that was sponsored by the former. The resulting network has 547 nodes and 33,937 edges; the largest strongly connected component of the network consists of 438 nodes and 28,194 edges that represent legislators in the House of Representatives.⁵

We use the legislative effectiveness score [76] computed for the corresponding sitting as the network’s external information access measure. The score itself quantifies each representative’s ability to advance their agenda through the legislative process, from introducing a bill to enacting it as a law, based on 15 weighted indicators [27].

Google Scholar Finally, we analyze a co-authorship network scraped from Google Scholar.⁶ As in the DBLP co-authorship network, each node represents a scholar and scholars are connected by an edge if they have co-authored a paper. The data were originally collected in May of 2015 by [19].⁷ They used a web crawler to search each letter in the English alphabet using the Google Scholar “search authors” feature and parsed those results to gather the scholars’ metadata and publication records. We added edges between these author nodes if two authors had the same paper listed in their set of published papers. This resulted in a total of 812,351 nodes and 262,933,633 edges, with 391,642 nodes and 104,647,630 edges in the largest connected component. We use total citation count as an external measure for influence (these citation counts are available for all nodes) as well as h -index (as recorded by Google Scholar).

We include the Google Scholar network in our analysis mostly due to its large size, the implications of which we will discuss further in Section 7. For now, we focus on the setup and analysis of the other networks.

5.2 Experimental setup

Calculating information access signatures We calculate the information access vector of node j , $s_\alpha^G(v_j) = (p_{1j}, \dots, p_{ij}, \dots, p_{nj})$, and estimate each p_{ij} by setting node i as the seed and running $S = 10,000$ independent cascade simulations. We then use the fraction of simulations in which node j received the information to estimate its information access probability.

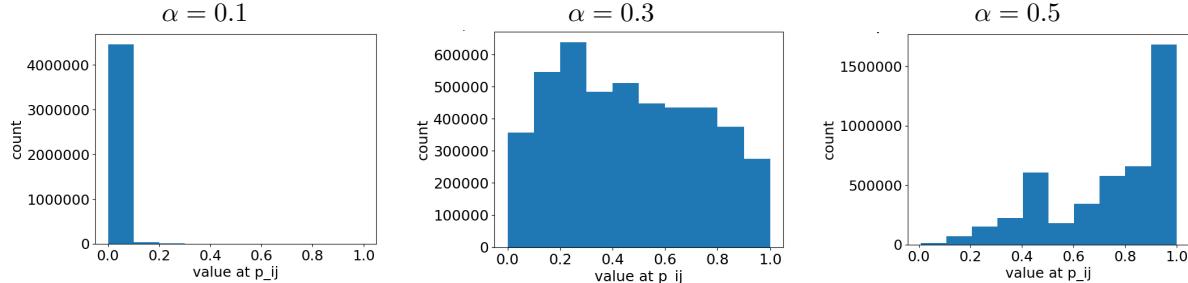


Figure 3: Histograms of probability values for the DBLP network and $\alpha = \{0.1, 0.3, 0.5\}$.

Choosing α The information access representation, and thus the clustering, are dependent on the choice of α . Since α represents the probability that information is transmitted along an edge, the choice of α is domain-dependent. We will thus present our results across a range of α values. However, choosing an appropriate range for α (which is also domain-dependent) is important in order to find p_{ij} values that are not zero or one: at the extremes based on a specific domain, most p_{ij} values in the information access signatures will be zero (indicating that the α value was too small to allow for information transmission across the network) or most will be one (indicating that the α value was large enough that essentially all

⁵There are more than 435 such legislators because of special elections held during the 114th sitting.

⁶<https://scholar.google.com>

⁷For the original paper, Chen et al. eliminated 409,961 nodes for having unreliable data. We use the full dataset, which includes any inaccuracies present in Google Scholar.

nodes received information from essentially all other nodes). Thus, information access signatures are most interesting when a range of probability values p_{ij} are included in the resulting information access signatures. To determine those ranges, we considered the histograms for each dataset (a selection of these are shown in Figure 3 with the full set of histograms in Appendix A.1. We choose $\alpha \in [0.1, 0.7]$ at increments of 0.1 for the DBLP and Twitch datasets and $\alpha \in [0.01, 0.07]$ at increments of 0.01 for the Co-sponsorship dataset since these ranges included a wide spread of histograms, including those with essentially all probabilities close to 0 and also where essentially all probabilities are close to 1. Larger values ranging up to the maximum of 1 were considered, but above the chosen range essentially all p_{ij} values were 1. In order to further assess whether the examined α values would provide enough range and granularity, the resulting clusterings were also examined for consistency across α values (see Appendix A.1.1).

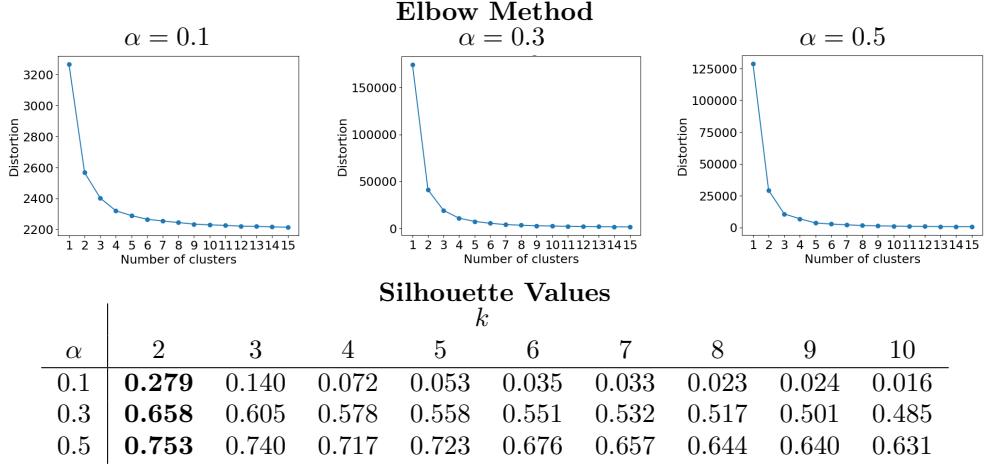


Figure 4: Elbow method charts and silhouette values for the DBLP dataset demonstrating that $k = 2$ for information access clustering on this dataset.

Choosing k Each clustering is dependent on the value of the number k of clusters chosen. This also depends on the domain / network. We use the silhouette value [65] and elbow method to determine the appropriate values of k for each dataset, looking for consistency across those two methods (see Figure 4 for selected α value results on the DBLP dataset and Appendix A.2 for full results). The DBLP and Co-sponsorship datasets have matching silhouette values and elbow method plots across all considered α values indicating that $k = 2$. The Twitch dataset elbow method plots indicate that $k = 2$ for all α values, while the silhouette values indicate that $k = 2$ for low values of α and has some slightly larger silhouette values for other k values at higher α values. Since the elbow method plots consistently indicate that $k = 2$, the silhouette value differences are small, and the Twitch dataset has a known domain-specific $k = 2$ value from the “partner” flag described above, we also use $k = 2$ for the Twitch dataset. We note here the coincidence that k was found to be 2 for all of our networks; though we do not explore in this paper potential reasons behind and implications of this.

5.3 Clusters and external information access measures

We hypothesize that since information access also controls and/or is correlated with people’s access to other resources, this implies that certain external measures such as academic productivity and privileged status in an affiliate program should correlate with access to information. We thus further hypothesize that *external information access measures will be different between different clusters*, indicating a difference in the social capital of each cluster and the structural similarity within clusters in terms of network position. In this section, we evaluate this hypothesis for the DBLP, Twitch, and congressional co-sponsorship networks. Specifically, each cluster induces a distribution of external information access measures. If these distributions significantly differ from one another, this provides evidence consistent with our hypothesis. We use

Clustering	DBLP			Twitch		Co-sponsorship Legislative Effectiveness
	Citation Count	PhD Rank	Job Rank	Partner	$\log(\text{Views})$	
Info Access	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$
Spectral	0.762	2.31e-3	1.30e-3	1	0.857	0.050
Fluid Comm.	0.0097	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$
Louvain	0.164	0.226	0.013	1	0.224	$< 10^{-7}$
Role2Vec	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$	0.007	$< 10^{-7}$	$< 10^{-7}$
Core-Periphery	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$	$< 10^{-7}$	0.472

Table 2: Kruskal-Wallis and Fisher Exact test (for the partner categorical data) p -values testing to see if the distributions of the external information access measures are different across clusters. Minimum p -values across all tested α values are given for the information access clusters, shown with an applied Bonferroni correction of 10. The fluid communities algorithm includes some randomness, so values shown are the minimum p -value for that attribute over 10 runs with a Bonferroni correction of 10 applied. Resulting p -values less than or equal to 0.05 are shown in bold. Very small p -values are shown as $< 10^{-7}$.

the Kruskal-Wallis and the Fisher exact tests to check the similarity of distributions. There are 10 total experiments for each dataset (by varying α across a range of plausible values), and so we apply a Bonferroni correction factor of 10 to all reported results⁸ See Table 2 for all results.

Looking at the probability density functions per cluster with respect to the external information measure allows us to better understand cluster composition. These results are shown for select α values on the DBLP dataset in Figure 5 with full results in Appendix B. We find that one of the DBLP clusters has more researchers with higher citation counts and better job and PhD institution ranks than the other, one of the Twitch clusters has more individuals with higher views and partner status, and one of the Co-sponsorship clusters has more individuals with higher legislative effectiveness scores. These clusters contain the more influential individuals in the network, with higher access to information and better ability to spread that information or reach important individuals in the network. Overall, these results thus show that information access clustering is able to create clusters that separate nodes by external information access measure across all three datasets. This evidence is consistent with our hypothesis and indicates that *information access representations encode information relevant to real-world social capital*.

6 Experiments Part II: Comparisons

We now consider our second question: whether other clustering methods create clusterings that separate nodes based on external information access measures, and whether information access clustering is consistently similar to one or more of these other methods. I.e. do existing methods already identify emergent network privilege by creating essentially the same clusters as those chosen by information access clustering? We will show that the answer is no.

6.1 Comparison methods

The clustering method we introduce does not cleanly belong to the major graph clustering method groups (see Section 2), so we compare against one method from each of the following groups:

Spectral methods. The well-known spectral clustering technique uses the eigenvectors of the graph Laplacian as node representation, which it then uses to cluster the vertices. We compare to `sklearn's SpectralClustering` [60] with default parameters: as many eigenvectors for the representation as the number of clusters, and a normalized Laplacian.

⁸This correction is 10 and not 7 since tests not included in the paper were originally run on a larger range of α values; we exclude three of these runs from the paper because of analysis described in Section 5.2.

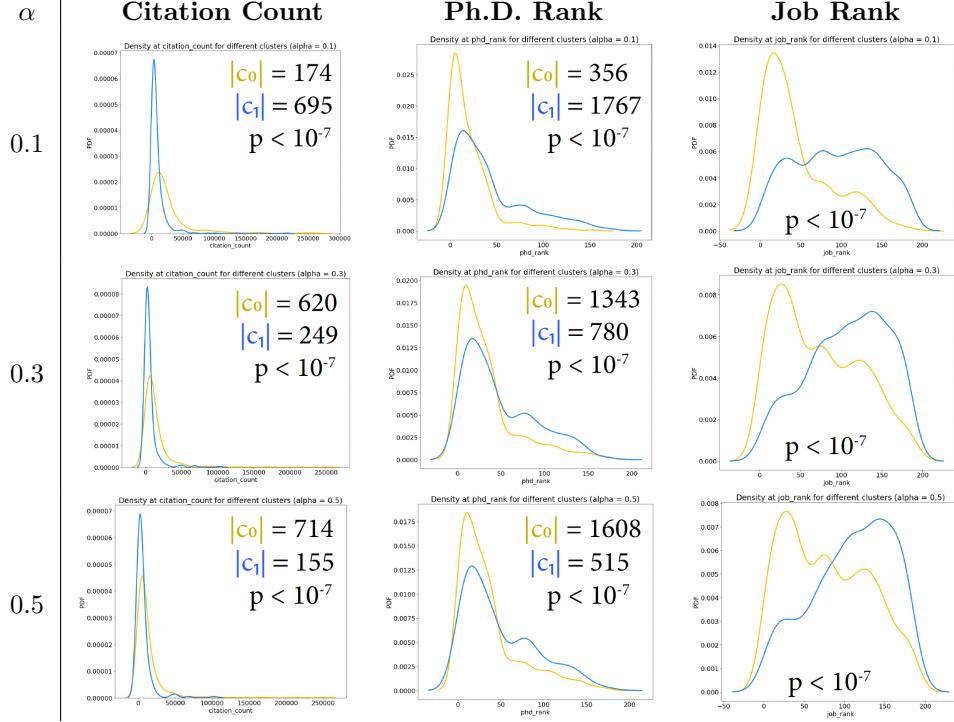


Figure 5: Information access clusterings on the DBLP dataset for $\alpha = \{0.1, 0.3, 0.5\}$. The probability density function (PDF) per cluster is shown with respect to the indicated external information access measure.

Community detection. Community detection methods are concerned with finding dense subgraphs based on direct connections between individuals. We compare to two such methods. The Louvain method attempts to optimize the modularity of a cluster (the relative density of links internal to a cluster versus those connecting outside the community [12]⁹). The fluid communities method is based on the idea of propagation of a fluid within a community, reaching an equilibrium with neighboring communities in the network from randomly chosen seed nodes [59].¹⁰

Role detection. Role detection algorithms cluster nodes together if they serve similar roles in the network. Nodes are often based on locally defined motifs, e.g., nodes that bridge two otherwise disconnected components. We compare to role2vec [3]¹¹. This method creates a vector representation of the network by considering random walks that balance an outward exploratory focus with returning to the start of the walk. Role2vec additionally takes an input motif to help guide these walks; we use an identity matrix motif to focus on individual nodes, since our other comparisons also assume no additional information.

Core-periphery clustering. Finally, core-periphery methods assume a network has a *core* group of nodes that are densely connected and a *periphery* that is well-connected to the core but not well connected within the periphery; a classic example is a star graph. We compare to Rombach et al.’s method [63].¹² This approach learns a mapping from nodes to the $[0, 1]$ interval with larger value signifying greater indication of being a core vertex. The mapping is constrained to agree with the core-periphery characterization: two periphery vertices should not be connected to each other, and two core vertices should.

⁹<https://python-louvain.readthedocs.io/en/latest/>

¹⁰[https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community. async_fluid.html](https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community._async_fluid.html)

¹¹<https://github.com/benedekrozemberczki/role2vec>

¹²<https://github.com/skojaku/core-periphery-detection/blob/master/cpnet/Rombach.py>

6.2 Assessing clustering similarity

Using the same experimental setup described in the previous section, we find (see Table 2) that many of these comparison methods create clusters that are separated in terms of their external information access measure.¹³ Overall, both community detection methods succeed at separating clusters according to these access measures for the co-sponsorship data, but Louvain fails on one more external access measures for the DBLP and Twitch data. Fluid community results differ drastically depending on the random initialization of the clusters; some resulting clusterings successfully distinguish information access measures while others do not. The core-periphery method, on the other hand, fails to separate based on the legislative effectiveness measure for the co-sponsorship data while it succeeds across all measures for the DBLP and Twitch data. The role detection method *does* successfully create clusters with different information access distributions across all datasets and access measures, as does information access clustering.

α	Spectral Clustering	Fluid Communities	Louvain	Role2Vec	Core-Periphery
0.1	~ 0	~ 0	~ 0	~ 0	0.73
0.3	~ 0	~ 0	~ 0	0.49	~ 0
0.5	~ 0	~ 0	0.01	0.38	~ 0

Table 3: For the DBLP dataset and each α -parameterized clustering, the above table gives the adjusted rand index indicating the difference between the resulting information access clustering and the indicated clustering method. The adjusted rand index is 0 when two clusterings do not agree on any pair of points. For legibility, adjusted rand index values less than 0.01 are shown as ~ 0 . Fluid communities method given values are the mean across 10 random runs.

To further assess the similarity of these clustering methods, we directly compared the resulting clusterings to determine whether nodes were assigned to the same groupings across clusterings (see Table 3 for selected results on the DBLP dataset and Appendix C for full results). No method created clusterings that were similar to information access clusterings across all datasets.

Inf. Acc. α	DBLP		Twitch		Co-sp.	
0.1 or 0.01	1	72	1	2461	1	1
0.2 or 0.02	1	374	1	2674	1	1
0.3 or 0.03	1	455	1	2169	1	1
0.4 or 0.04	1	422	1	2012	1	1
0.5 or 0.05	1	414	1	2015	1	1
0.6 or 0.06	1	233	1	1212	1	11
0.7 or 0.07	1	224	1	1173	1	6
Spectral	1	1	1	5	1	1
Fluid C*.	3	4	45	48	1	1
Louvain	1	1	1	1	1	1
Role2Vec	4	299	253	674	1	1
Core-Peri.	1	102	1	1756	1	1

Table 4: Number of connected components per cluster by clustering type and dataset. For clustering methods that distinguish between more or less influential clusters along at least one external information access measure (see Table 2), the more influential cluster is listed first and the entries are given in bold. Fluid communities numbers are the mean values over different seeds.

To further investigate and understand the similarities and differences between these methods, we deter-

¹³In some cases where they don't, the clusterings produced with $k = 2$ on some datasets have a very small number of nodes in one of the clusters. In these cases, sample sizes are so small that they provide no statistical evidence of the difference between cluster distributions. We observe this for spectral clustering and the Louvain method, on both the DBLP and Twitch datasets.

mined the number of connected components per cluster for each of these methods (see Table 4). This allows us to see that information access clustering (for $k = 2$) creates one cluster with a single connected component and another cluster that is a collection of disconnected components on the DBLP and Twitch datasets. In all cases where these clusters distinguish the external access measure, the cluster with the single connected component also contains more of the higher information access (privileged) nodes. In this way, information access clustering is similar to the core-periphery methods that create a single core or influential cluster that’s connected and another cluster to represent the periphery. Again, on the co-sponsorship dataset we see that information access clustering is more similar to community detection methods, with both clusterings creating two single component clusters (for all significant α values). Finally, we see that role2vec does not follow the same core-periphery pattern as information access clustering for the DBLP and Twitch data, creating two clusters with multiple components. Thus, information access clustering appears to act differently from all examined methods and is able to identify privileged clusters across both core-periphery and community structured networks.

7 Experiments Part III: Large networks

The previous experiments considered networks that have at most 7,126 nodes. The Google Scholar network, however, has 391,642 nodes and 104,647,630 edges in its largest strongly connected component which makes calculating the full information access signature for each node prohibitively slow. In this section, we assess selection of a smaller signature based on a number of possible seed selection methods. We first use the DBLP co-authorship network to assess the extent to which the resulting information access clustering matches that performed on a full representation, and then consider the Google Scholar co-authorship network.

7.1 Computing information access representations for large graphs

Computing the information access signature of a node $v_j \in V$ demands determining its probability of receiving information from every other node in V . This becomes computationally intensive for large networks. We therefore consider a smaller signature (abusing notation slightly), $s_\alpha^I(v_j) = (p_{1j}, \dots, p_{ij}, \dots, p_{bj})$ for all $v_i \in I$, where $I \subset V$ and $|I| = b$, i.e. instead of considering the signature representing each other node as a possible source of information, we choose a subset of *seed* nodes and create each information access signature only in terms of those nodes.

Based on these smaller signatures, we determine the representation $R_\alpha^I = \{s_\alpha^I(v_i) | v_i \in V\}$ and treat this as the network representation. This creates a matrix representation of size n by b , i.e., this smaller information access signature is created for each node. In cases where the graph is too large for local memory, it’s not possible to update all signatures simultaneously. However, the probabilities $\{p_{i1}, \dots, p_{ij}, \dots, p_{in}\}$ for each $v_i \in I$ can be computed independently based on a single information propagation experiment and then merged to create the representation. Since we use the independent cascade method for our experiments, we do this by simulating an independent cascade starting from the single node v_i and determining reachability for all nodes in the graph. As described in Section 5.2, these are run 10,000 times and averaged to determine the probability estimates. Usefully for large networks, computing the probabilities one seed at a time in this way will likely mean that a large portion of the graph remains unreached by the independent cascade, and thus does not need to be stored.

Ideally, the smaller access signatures resulting from the seed set I would be similar to those resulting from the full network, and would also result in similar clusterings. We leave a full theoretical analysis of how to select such an I as an open problem; and limit ourselves to four heuristic approaches to picking I . These seed selection heuristics are designed to take advantage of intuition that choosing highly connected or otherwise central nodes as seeds will allow most nodes in the graph to be reached and thus allow those, potentially peripheral, nodes to have non-zero terms in the resulting signature. Thus we will evaluate three deterministic strategies that rank nodes according to a centrality criterion and pick the top b of these. The three criteria we use are (i) (global) PageRank, (ii) betweenness centrality, and (iii) node degree (also known as degree centrality). Additionally, we examine the efficacy of choosing b seeds uniformly at random.

With these modifications, the time to compute the full representation can be reduced from $O(Smn)$ to $O(Smb)$. For each seed we compute the signature independently (which takes $O(Sm)$ by doing S repetitions of breadth first search from that seed). Therefore, computing this smaller representation takes $O(C) + O(S \cdot$

$m \cdot b$) time and $O(bn)$ space where S is the number of simulations we run to estimate the probabilities, C is the time it takes the chosen selection strategy to select all seeds, and n and m are, respectively, the number of nodes and edges in the graph. In the case of the random seed selection strategy, this is simply $O(Smb)$.

7.2 Experimental results

Given the goal of using the seed set I to compute a smaller information access representation that still generates clusterings that are similar to those from the full representation, we next consider the DBLP network. The DBLP network is small enough to compute the full representation, so we compare its information access clustering on the smaller signatures and on the full representation, using the adjusted rand index (adjusted rand index values are 1 when two clusterings agree on all pairs of points). We consider the four selection strategies for choosing the set of seed nodes (see Section 7.1) and run the information access clustering on the resulting smaller representation.

We find that clusterings that are very similar to those developed via the full representation can be created even with a small number of selected seeds. We tested seed sets of size 5 to 70 at increments of 5 and found that for $\alpha \in [0.2, 0.7]$, the adjusted rand index was at least 0.99 for the clusterings resulting from all centrality-based seed selection strategies; randomly selected seeds generated clusterings with adjusted rand indexes at least 0.97 for the same parameters, so the random strategy was also highly successful. While the DBLP dataset has 2,123 nodes, smaller signatures with only 10 or more seeds create clusterings with an adjusted rand index of at least 0.97 when compared with information access clustering for values of $\alpha > 0.1$ over all selection strategies. We find that all three centrality-based selection strategies perform better than random selection, but random selection also performs well (see results for $\alpha = 0.1$ and $\alpha = 0.2$ in Figure 6 and full results in Appendix Table 7), creating clusterings that are essentially the same as those from the full representation. Since random sampling is faster than the other strategies, we adopt that strategy for experiments on the larger Google Scholar network.

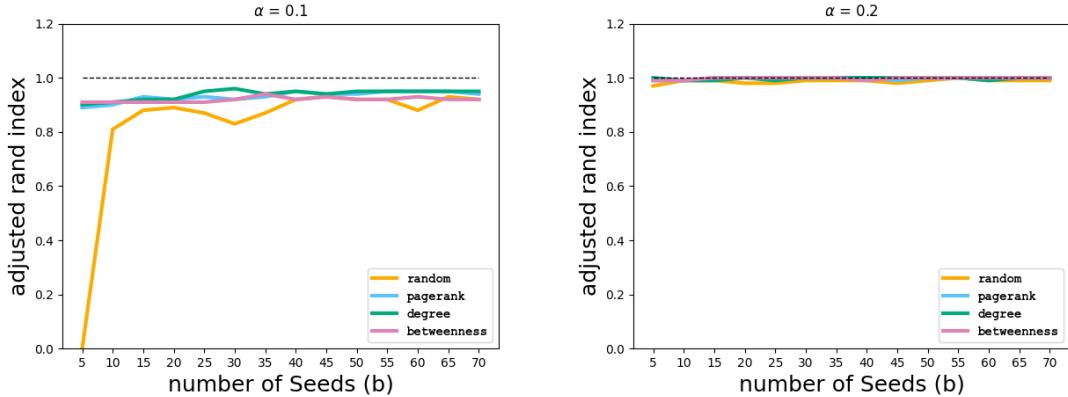


Figure 6: Results for four seed selection strategies to create smaller information access representations are shown for the DBLP dataset. Adjusted rand indices of the resulting clusterings when compared to the full representation clustering are shown for $\alpha = 0.1$ (left) and $\alpha = 0.2$ (right) with seed set sizes at increments of 5 from 5 to 70. The results for $\alpha > 0.2$ are essentially the same as those for $\alpha = 0.2$ and are given in Appendix Table 7. Adjusted rand indices are 1 when two clusterings are the same.

Recall from Section 5.1 that the Google Scholar network has 391,642 nodes. We choose the size of I to be $b = 632$ since that is the square root of the total number of nodes in the network. This simple heuristic for sample size selection was found to be an effective size choice in the previous section on the DBLP dataset which had 2,123 nodes (where a sample of 46 nodes produced adjusted rand indices of at least 0.99 on most α values indicating essentially the same clustering as would have been calculated using the full information access representation). Using the same experimental setup as described in Section 5.2, we considered the histograms showing the prevalence of varying p_{ij} values within the information access signatures varying over $\alpha \in \{0.01, 0.03, 0.05, 0.4\}$. The resulting range of probabilities includes α values that generate low, medium,

and high spreading scenarios, as desired. The silhouette values and elbow method plots were generated for the chosen α values, and all methods determined that $k = 2$. Full results of these experimental setup experiments can be found in Appendix D.1.

Information access signatures were calculated on a shared computing cluster, with run time depending on the availability of node resources as well as α . On average the run time for a single seed node was roughly 5 minutes, for a total of approximately 2 days to generate the information access representation for a specific α given the seed size of 632. As before, we investigate the correlation between clusterings and the Google Scholar network’s external information access measure. We use the Kruskal-Wallis test and find evidence ($p < 10^{-7}$) across all α values that the distribution of citation counts and h -index values is different between clusters in the information access clusterings.

8 Discussion and Conclusion

In this paper, we introduced a representation of individuals in a network based on their information access, creating a mathematical representation of social capital in a network. Given this social capital representation, we introduced a technique to cluster individuals based on network privilege, as indicated by information access. We provided both social theory and mathematical formal insights into what the underlying representation captures, showed that calculating this representation is $\#P$ -hard, and provided practical heuristics for its calculation on large graphs. Using real-world data, we validated the encoding of social capital information in the introduced information access representations, demonstrating that clustering on this representation effectively separates individuals based on external measures of information access. We showed experimentally that these clusterings are different than existing community-based, role detection, and core-periphery methods.

While we chose these information access measures to be purposefully *external* to the network, given that the information access clustering was able to create clusters that distinguish these measures using only information encoded in the network, these measures are not truly external. This brings up an interesting question of causality – did the network structure impact these “external” measures or did the measures of social capital impact the structure of the network? We leave this interesting, potentially domain-specific, question for future work.

References

- [1] Frédéric Adam. 2008. Using Network Analysis for Understanding How Decisions are Made. In *Encyclopedia of Decision Making and Decision Support Technologies*. IGI Global, 950–957.
- [2] Charu C. Aggarwal and Haixun Wang. 2010. *A Survey of Clustering Algorithms for Graph Data*. Springer US, 275–301.
- [3] Nesreen K Ahmed, Ryan A Rossi, John Boaz Lee, Xiangnan Kong, Theodore L Willke, Rong Zhou, and Hoda Eldardiry. 2018. Learning Role-based Graph Embeddings. *stat* 1050 (2018), 7.
- [4] Junaid Ali, Mahmoudreza Babaei, Abhijnan Chakraborty, Baharan Mirzasoleiman, Krishna P Gummadi, and Adish Singla. 2019. On the fairness of time-critical influence maximization in social networks. *arXiv preprint arXiv:1905.06618* (2019).
- [5] Uri Alon and Eran Yahav. 2021. On the Bottleneck of Graph Neural Networks and its Practical Implications. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=i800Ph0CVH2>
- [6] Adrian Arnaiz-Rodriguez, Ahmed Begga, Francisco Escolano, and Nuria Oliver. 2022. DiffWire: Inductive Graph Rewiring via the Lovász Bound. *arXiv:2206.07369 [cs.LG]*
- [7] Mark Balnaves. 1993. The sociology of information. *The Australian and New Zealand journal of sociology* 29, 1 (1993), 93–111.
- [8] Ashkan Bashardoust, Sorelle Friedler, Carlos Scheidegger, Blair D. Sullivan, and Suresh Venkatasubramanian. 2023. Reducing Access Disparities in Networks Using Edge Augmentation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) (FAccT ’23). Association for Computing Machinery, New York, NY, USA, 1635–1651. <https://doi.org/10.1145/3593013.3594105>
- [9] Ruben Becker, Gianlorenzo D’Angelo, and Sajjad Ghobadi. 2023. Improving Fairness in Information Exposure by Adding Links. *arXiv:2302.13112 [cs.SI]*
- [10] Ruben Becker, Gianlorenzo D’Angelo, Sajjad Ghobadi, and Hugo Gilbert. 2021. Fairness in Influence Maximization through Randomization. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 14684–14692. <https://ojs.aaai.org/index.php/AAAI/article/view/17725>
- [11] Hannah C. Beilinson, Nasanbayar Ulzii-Orshikh, Ashkan Bashardoust, Sorelle A. Friedler, Carlos E. Scheidegger, and Suresh Venkatasubramanian. 2020. Clustering via Information Access in a Network. *arXiv:2010.12611 [cs.SI]*
- [12] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.
- [13] Stephen P Borgatti and Martin G Everett. 2000. Models of core/periphery structures. *Social networks* 21, 4 (2000), 375–395.
- [14] danah boyd, Karen Levy, and Alice Marwick. 2014. The Networked Nature of Algorithmic Discrimination. *Data & Discrimination: Collected Essays* (2014), 43–57.
- [15] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. 2007. On modularity clustering. *IEEE transactions on knowledge and data engineering* 20, 2 (2007), 172–188.
- [16] Ronald S. Burt. 2004. Structural Holes and Good Ideas. *Amer. J. Sociology* 110, 2 (2004), 349–399.
- [17] Ronald S Burt. 1987. Social contagion and innovation: Cohesion versus structural equivalence. *American journal of Sociology* 92, 6 (1987), 1287–1335.

- [18] Ronald S. Burt. 2000. The Network Structure Of Social Capital. *Research in Organizational Behavior* 22 (2000), 345–423.
- [19] Yang Chen, Cong Ding, Jiayao Hu, Ruichuan Chen, Pan Hui, and Xiaoming Fu. 2017. Building and Analyzing a Global Co-Authorship Network Using Google Scholar Data. In *WWW*. 1219–1224.
- [20] Tristan Claridge. 2017. How to measure social capital. *Social Capital Research and Training* (2017).
- [21] Tristan Claridge. 2018. What is structural social capital? *Social Capital Research and Training* (2018).
- [22] Tristan Claridge. 2022. The difference between social capital and cultural capital. *Social Capital Research and Training* (2022).
- [23] Aaron Clauset, Samuel Arbesman, and Daniel B. Larremore. 2015. Systematic inequality and hierarchy in faculty hiring networks. *Science Advances* 1, 1 (2015).
- [24] James S. Coleman. 1988. Social Capital in the Creation of Human Capital. *Amer. J. Sociology* 94 (1988), S95–S120.
- [25] James S. Coleman, Elihu Katz, and Herbert Menzel. 1966. *Medical Innovation: A diffusion study*. Bobbs-Merril, New York.
- [26] Peter Csermely, András London, Ling-Yun Wu, and Brian Uzzi. 2013. Structure and dynamics of core/periphery networks. *Journal of Complex Networks* 1, 2 (2013), 93–123.
- [27] Lawrence C. Dodd and Bruce I. Oppenheimer. 2017. *Congress Reconsidered*. CQ Press; Eleventh Edition. 248–250 pages.
- [28] Pedro Domingos and Matt Richardson. 2001. Mining the Network Value of Customers. *KDD* (11 2001).
- [29] Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. 2022. EDITS: Modeling and Mitigating Data Bias for Graph Neural Networks. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) (*WWW '22*). Association for Computing Machinery, New York, NY, USA, 1259–1269. <https://doi.org/10.1145/3485447.3512173>
- [30] Yushun Dong, Song Wang, Yu Wang, Tyler Derr, and Jundong Li. 2022. On Structural Explanation of Bias in Graph Neural Networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) (*KDD '22*). Association for Computing Machinery, New York, NY, USA, 316–326. <https://doi.org/10.1145/3534678.3539319>
- [31] Benjamin Fish, Ashkan Bashardoust, danah boyd, Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Gaps in Information Access in Social Networks. In *WWW*. 480–490.
- [32] Santo Fortunato. 2010. Community detection in graphs. *Physics reports* 486, 3-5 (2010), 75–174.
- [33] James H. Fowler, Andrew Scott Waugh, and Yunkyu Sohn. 2017. Cosponsorship Network Data. <http://jhfowler.ucsd.edu/cosponsorship.htm>.
- [34] Shayan Oveis Gharan and Luca Trevisan. 2014. Partitioning into expanders. In *SODA*. 1256–1266.
- [35] Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Lio', and Michael Bronstein. 2023. On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology. arXiv:2302.02941 [cs.LG]
- [36] Mark Granovetter. 1978. Threshold models of collective behavior. *American journal of sociology* 83, 6 (1978), 1420–1443.
- [37] Mark S. Granovetter. 1973. The Strength of Weak Ties. *The American Journal of Sociology* 78, 6 (1973), 1360–1380.

- [38] Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, and Lei Li. 2012. RolX: Structural Role Extraction & Mining in Large Graphs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Beijing, China) (*KDD '12*). Association for Computing Machinery, New York, NY, USA, 1231–1239. <https://doi.org/10.1145/2339530.2339723>
- [39] Keith Henderson, Brian Gallagher, Lei Li, Leman Akoglu, Tina Eliassi-Rad, Hanghang Tong, and Christos Faloutsos. 2011. It’s Who You Know: Graph Mining Using Recursive Structural Features. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Diego, California, USA) (*KDD '11*). Association for Computing Machinery, New York, NY, USA, 663–671. <https://doi.org/10.1145/2020408.2020512>
- [40] Herbert W Hethcote. 2000. The mathematics of infectious diseases. *SIAM review* 42, 4 (2000), 599–653.
- [41] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. *Social Networks* 5, 2 (1983), 109–137. [https://doi.org/10.1016/0378-8733\(83\)90021-7](https://doi.org/10.1016/0378-8733(83)90021-7)
- [42] Marleen Huysman and Volker Wulf. 2004. Social capital and information technology: Current debates and research. *Social capital and information technology* (2004), 1–16.
- [43] Matthew Jackson. 2019. *The Human Network: How Your Social Position Determines Your Power, Beliefs, and Behaviors*. Knopf Doubleday Publishing Group.
- [44] Zeinab S Jalali, Qilan Chen, Shwetha M Srikanta, Weixiang Wang, Myunghwan Kim, Hema Raghavan, and Sucheta Soundarajan. 2022. Fairness of Information Flow in Social Networks. *ACM Transactions on Knowledge Discovery from Data* (2022).
- [45] Zeinab S Jalali, Weixiang Wang, Myunghwan Kim, Hema Raghavan, and Sucheta Soundarajan. 2020. On the Information Unfairness of Social Networks. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 613–521.
- [46] Byron Kaldis. 2013. *Encyclopedia of philosophy and the social sciences*. Vol. 1. Sage.
- [47] Ravi Kannan, Santosh Vempala, and Adrian Vetta. 2004. On Clusterings: Good, Bad and Spectral. *J. ACM* 51, 3 (May 2004), 497–515.
- [48] Kedar Karhadkar, Pradeep Kr. Banerjee, and Guido Montufar. 2023. FoSR: First-order spectral rewiring for addressing oversquashing in GNNs. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=3YjQfCLdrzz>
- [49] Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2018. Homophily influences ranking of minorities in social networks. *Scientific reports* 8, 1 (2018), 1–12.
- [50] David Kempe, Jon M. Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *KDD*. 137–146.
- [51] Eun Lee, Fariba Karimi, Claudia Wagner, Hang-Hyun Jo, Markus Strohmaier, and Mirta Galesic. 2019. Homophily and minority-group size explain perception biases in social networks. *Nature human behaviour* 3, 10 (2019), 1078–1087.
- [52] Yuchen Li, Ju Fan, Yanhao Wang, and Kian-Lee Tan. 2018. Influence maximization on social graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (2018), 1852–1872.
- [53] Wenlin Liu, Anupreet Sidhu, Amanda M Beacom, and Thomas W Valente. 2017. Social network theory. *The international encyclopedia of media effects* (2017), 1–12.
- [54] Anay Mehrotra, Jeff Sachs, and L Elisa Celis. 2022. Revisiting Group Fairness Metrics: The Effect of Networks. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–29.

- [55] Piotr Mikiewicz. 2021. Social capital and education—An attempt to synthesize conceptualization arising from various theoretical origins. *Cogent Education* 8, 1 (2021), 1907956.
- [56] Allison C Morgan, Dimitrios J Economou, Samuel F Way, and Aaron Clauset. 2018. Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science* 7, 1 (2018), 40.
- [57] Adam Mosseri. 2021. Shedding more light on how Instagram works. <https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works>
- [58] Marcos Oliveira, Fariba Karimi, Maria Zens, Johann Schaible, Mathieu Génois, and Markus Strohmaier. 2021. Mixing dynamics and group imbalance lead to degree inequality in face-to-face interaction. *arXiv preprint arXiv:2106.11688* (2021).
- [59] Ferran Parés, Dario Garcia Gasulla, Armand Vilalta, Jonatan Moreno, Eduard Ayguadé, Jesús Labarta, Ulises Cortés, and Toyotaro Suzumura. 2017. Fluid communities: A competitive, scalable and diverse community detection algorithm. In *International Conference on Complex Networks and their Applications*. Springer, 229–240.
- [60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [61] Joel M. Podolny and James N. Baron. 1997. Resources and Relationships: Social Networks and Mobility in the Workplace. 62, 5 (1997), 673–693. <https://doi.org/10.2307/2657354>
- [62] Aida Rahmattalabi, Shahin Jabbari, Himabindu Lakkaraju, Phebe Vayanos, Eric Rice, and Milind Tambe. 2020. Fair Influence Maximization: A Welfare Optimization Approach. *CoRR* abs/2006.07906 (2020). arXiv:2006.07906 <https://arxiv.org/abs/2006.07906>
- [63] M Puck Rombach, Mason A Porter, James H Fowler, and Peter J Mucha. 2014. Core-periphery structure in networks. *SIAM Journal on Applied mathematics* 74, 1 (2014), 167–190.
- [64] Ryan A. Rossi, Di Jin, Sungchul Kim, Nesreen K. Ahmed, Danai Koutra, and John Boaz Lee. 2020. On Proximity and Structural Role-based Embeddings in Networks: Misconceptions, Techniques, and Applications. In *Transactions on Knowledge Discovery from Data (TKDD)*. 36.
- [65] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.
- [66] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. 2019. Multi-scale Attributed Node Embedding. *arXiv: 1909.13021 cs.LG* (2019).
- [67] Michael Shapiro and Edgar Delgado-Eckert. 2012. Finding the probability of infection in an SIR network is NP-Hard. *Mathematical Biosciences* 240, 2 (2012), 77–84. <https://doi.org/10.1016/j.mbs.2012.07.002>
- [68] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22, 8 (2000), 888–905.
- [69] Martti Siisiainen. 2003. Two concepts of social capital: Bourdieu vs. Putnam. *International journal of contemporary sociology* 40, 2 (2003), 183–204.
- [70] Ana-Andreea Stoica and Augustin Chaintreau. 2019. Fairness in Social Influence Maximization. In *WWW*. 569–574.
- [71] Ian P. Swift, Sana Ebrahimi, Azade Nova, and Abolfazl Asudeh. 2022. Maximizing Fair Content Spread via Edge Suggestion in Social Networks. *Proc. VLDB Endow.* 15, 11 (sep 2022), 2692–2705. <https://doi.org/10.14778/3551793.3551824>

- [72] Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. 2022. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=7UmjRGzp-A>
- [73] Alan Tsang, Bryan Wilder, Eric Rice, Milind Tambe, and Yair Zick. 2019. Group-fairness in influence maximization. In *Proc. of the Int'l Joint Conf. on Artificial Intelligence*. AAAI Press, 5997–6005.
- [74] Leslie G Valiant. 1979. The complexity of enumeration and reliability problems. *SIAM J. Comput.* 8, 3 (1979), 410–421.
- [75] Ameya Velingker, Ali Kemal Sinop, Ira Ktena, Petar Veličković, and Sreenivas Gollapudi. 2022. Affinity-Aware Graph Networks. arXiv:2206.11941 [cs.LG]
- [76] Craig Volden and Alan Wiseman. 2020. Legislative Effectiveness Data. <https://thelawmakers.org/data-download>.
- [77] Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing* 17, 4 (2007), 395–416.
- [78] Xindi Wang, Onur Varol, and Tina Eliassi-Rad. 2021. Information Access Equality on Network Generative Models. *arXiv preprint arXiv:2107.02263* (2021).
- [79] Samuel F Way, Daniel B Larremore, and Aaron Clauset. 2016. Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks. In *WWW*. 1169–1179.
- [80] Yiguang Zhang, Jessy Xinyi Han, Illica Mahajan, Priyanjana Bengani, and Augustin Chaintreau. 2021. Chasm in Hegemony: Explaining and Reproducing Disparities in Homophilous Networks. *Proc. ACM Meas. Anal. Comput. Syst.* 5, 2 (2021), 16:1–16:38. <https://doi.org/10.1145/3460083>

A Experimental setup details

A.1 Choosing α

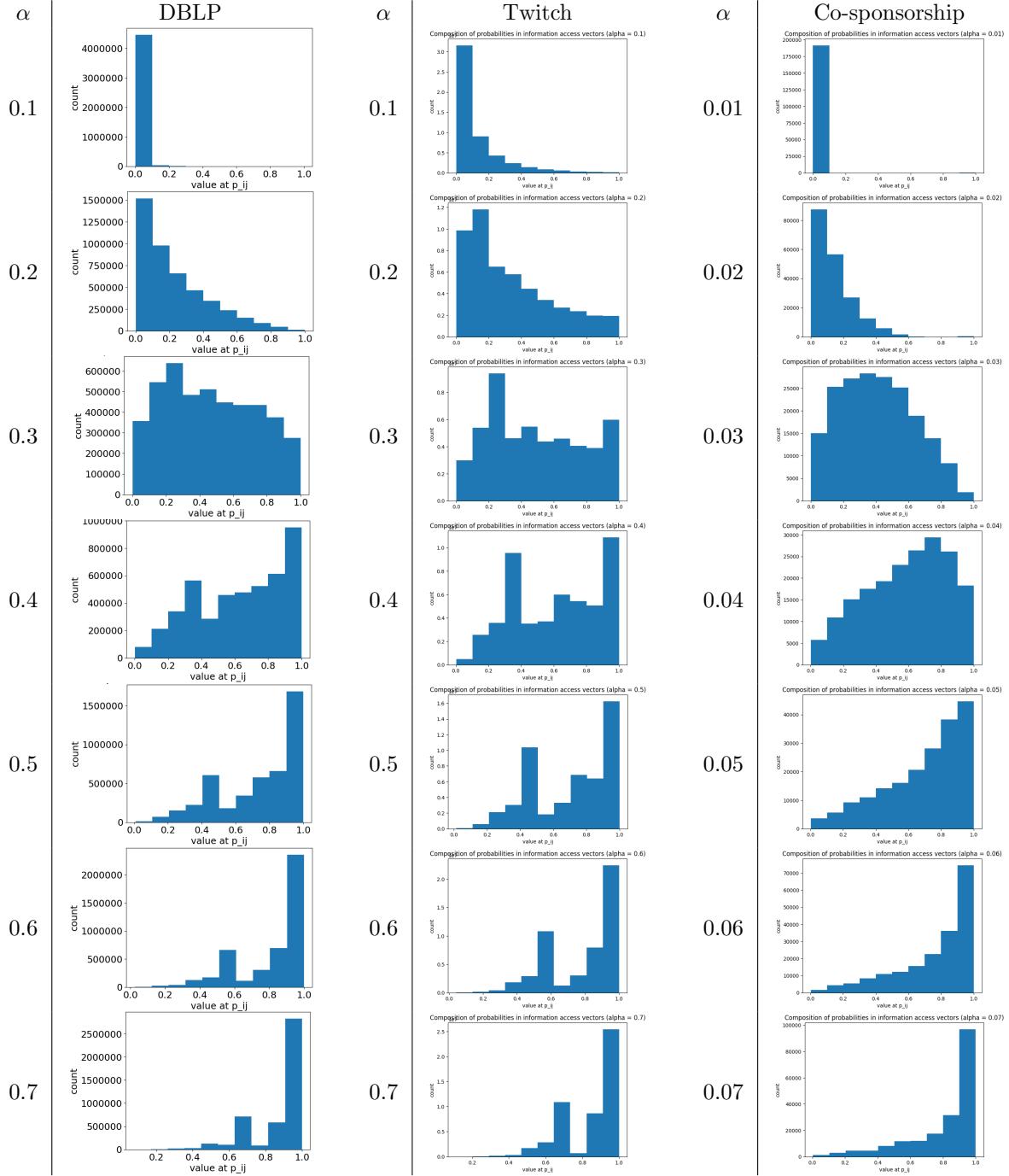


Figure 7: Histograms showing the prevalence of p_{ij} values within the information access signatures for the DBLP, Twitch, and Co-sponsorship data based on α value.

A.1.1 Clustering consistency across α ranges

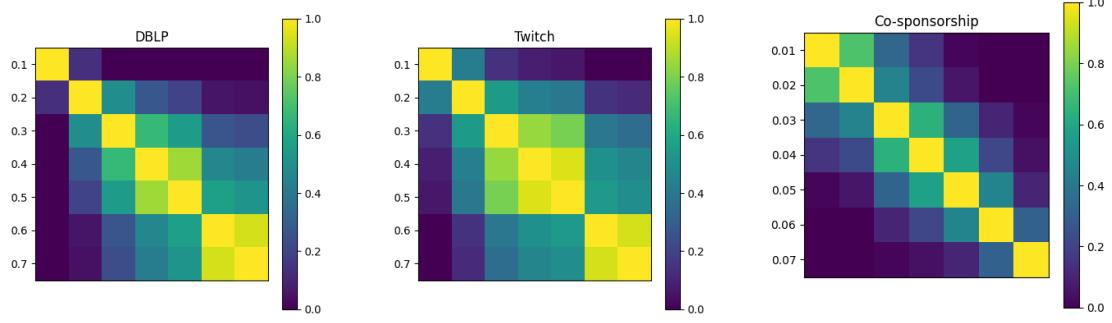


Figure 8: A visualization of the consistency of clusterings across the ranges of α investigated for each dataset. Each visualization shows the Adjusted Rand Index (ARI). We round ARIs up to 0 when negative to ensure consistency across colorscales for the three datasets.

As an additional validation step, here we investigate the extent to which clusterings generated by the different α values are consistent across the α values investigated. We do not expect consistent clusterings whenever the probability vectors are significantly different from one another, but expect consistency between clusterings with similar α values. Figure 8 shows the adjusted rand index values computed between the clustering assignments for all α values. Notice how for both the Twitch and DBLP datasets, there exists a clear range of α ($[0.2, 0.5]$ and $[0.3, 0.5]$ respectively) where the clusterings largely agree with one another. The situation is not as clear for the co-sponsorship dataset, but there, we can still see that similar α values tend to generate similar clustering assignments. Importantly, given the gradations along the diagonal it seems likely from this analysis that we are not missing important clusterings due to the choice of granularity of α values considered.

A.2 Choosing k

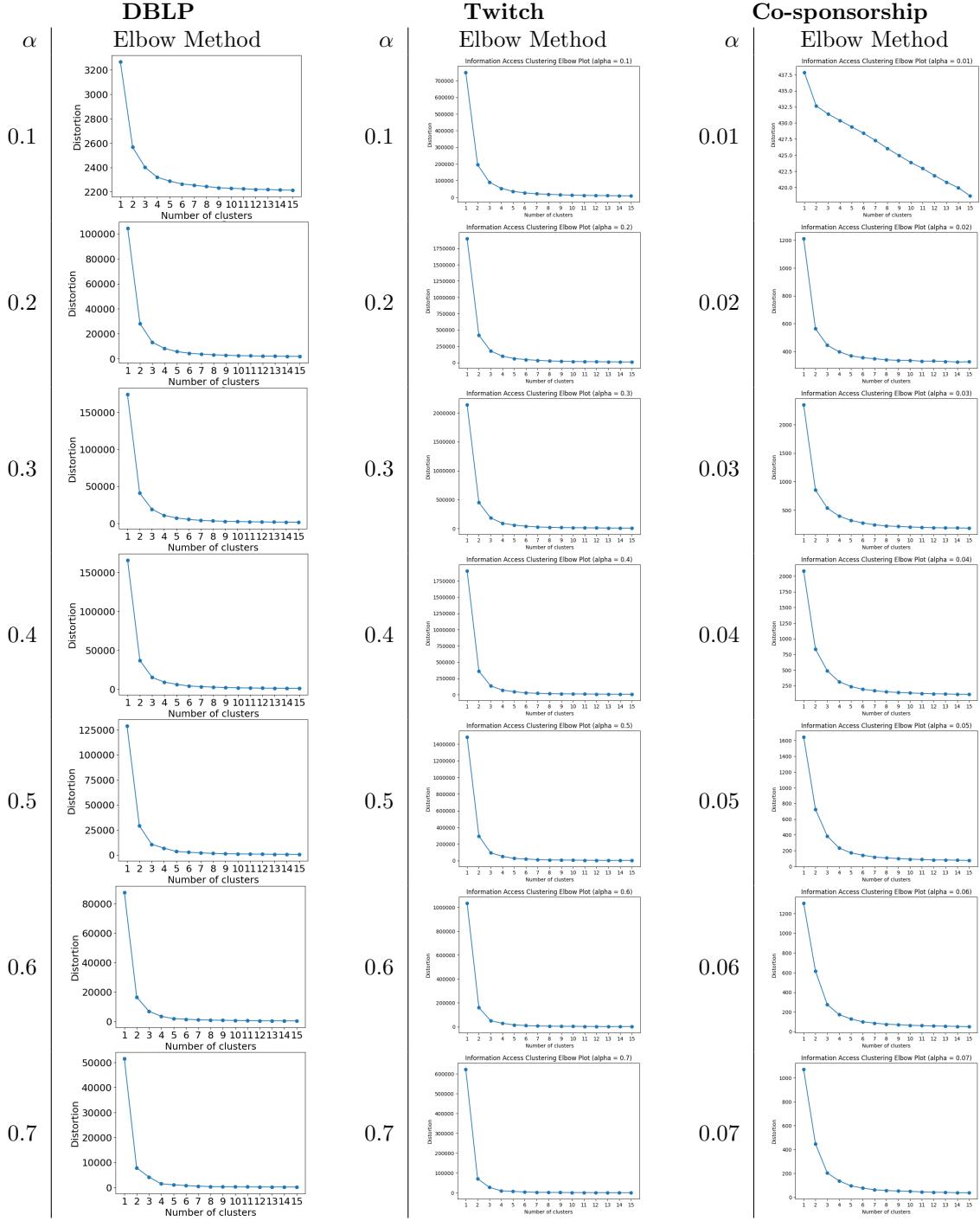


Figure 9: Elbow method plots for information access clustering on the DBLP, Twitch, and Co-sponsorship datasets.

		DBLP									
		k									
α		2	3	4	5	6	7	8	9	10	
0.1		0.279	0.140	0.072	0.053	0.035	0.033	0.023	0.024	0.016	
0.2		0.598	0.551	0.503	0.479	0.449	0.411	0.389	0.374	0.355	
0.3		0.658	0.605	0.578	0.558	0.551	0.532	0.517	0.501	0.485	
0.4		0.714	0.692	0.651	0.646	0.621	0.605	0.596	0.591	0.581	
0.5		0.753	0.740	0.717	0.723	0.676	0.657	0.644	0.640	0.631	
0.6		0.828	0.787	0.796	0.783	0.778	0.772	0.728	0.726	0.721	
0.7		0.876	0.836	0.847	0.823	0.829	0.820	0.821	0.821	0.774	

		Twitch									
		k									
α		2	3	4	5	6	7	8	9	10	
0.1		0.640	0.582	0.538	0.510	0.484	0.462	0.432	0.409	0.404	
0.2		0.652	0.620	0.600	0.594	0.581	0.571	0.569	0.564	0.544	
0.3		0.682	0.667	0.671	0.665	0.668	0.665	0.656	0.641	0.620	
0.4		0.732	0.730	0.729	0.733	0.734	0.728	0.721	0.720	0.711	
0.5		0.753	0.764	0.763	0.767	0.765	0.769	0.759	0.752	0.751	
0.6		0.822	0.823	0.828	0.815	0.817	0.816	0.805	0.793	0.802	
0.7		0.872	0.864	0.873	0.852	0.860	0.850	0.851	0.830	0.811	

		Co-sponsorship									
		k									
α		2	3	4	5	6	7	8	9	10	
0.01		0.009	0.004	0.002	0.0007	-0.0003	0.0003	0.00003	0.0001	0.00002	
0.02		0.411	0.268	0.188	0.153	0.125	0.096	0.081	0.074	0.061	
0.03		0.538	0.446	0.397	0.377	0.308	0.290	0.260	0.254	0.216	
0.04		0.584	0.543	0.498	0.429	0.397	0.388	0.335	0.272	0.266	
0.05		0.658	0.613	0.550	0.497	0.490	0.401	0.381	0.320	0.317	
0.06		0.738	0.664	0.584	0.554	0.499	0.508	0.417	0.396	0.377	
0.07		0.899	0.721	0.600	0.588	0.500	0.500	0.475	0.378	0.437	

Table 5: Silhouette values for varying α and k for the DBLP, Twitch, Co-sponsorship, and Google Scholar datasets for information access clustering. The largest silhouette value for each α is shown in bold.

B Clusters and external information access measures

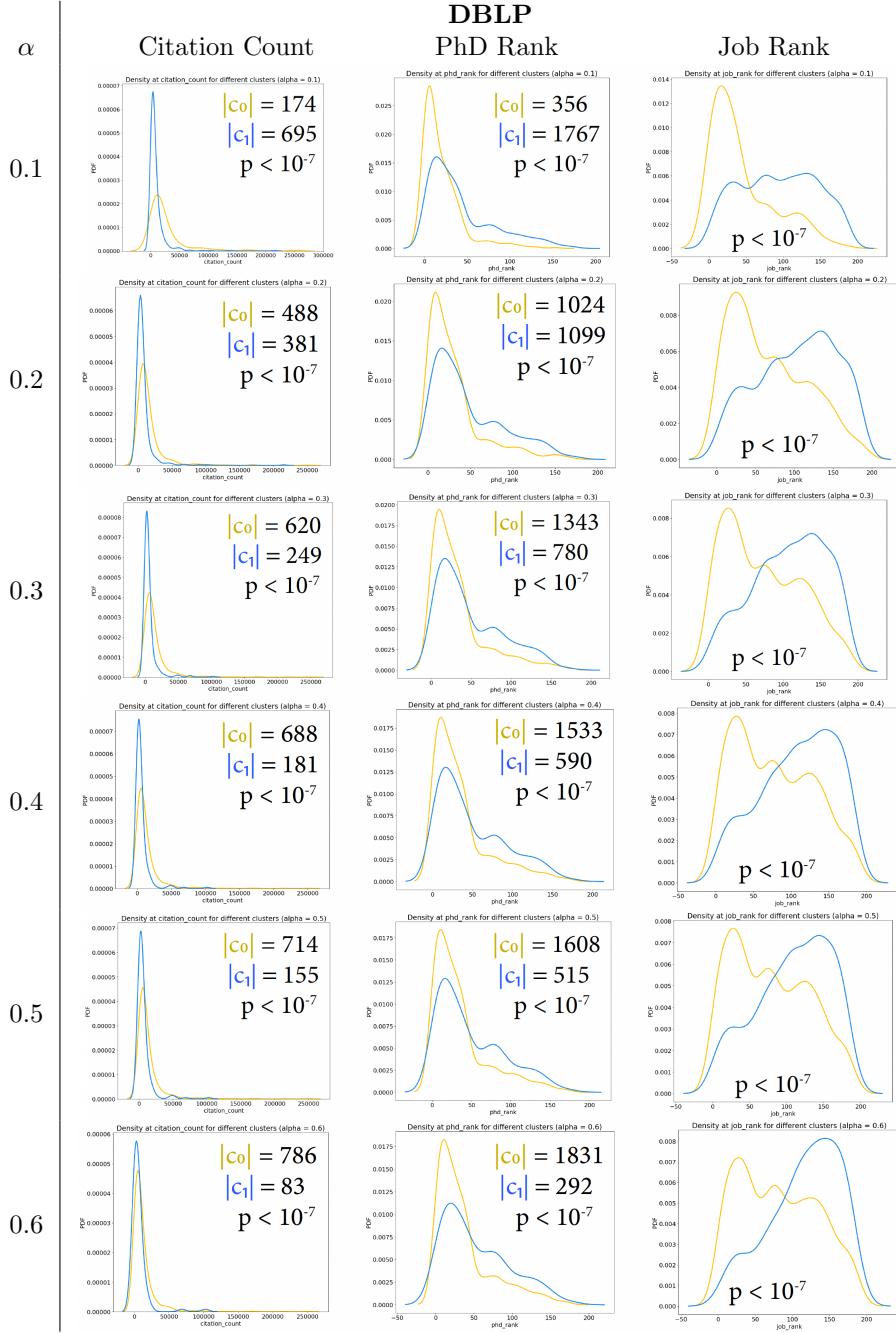


Figure 10: Results above are shown for the DBLP dataset. Clusterings are given for α values as shown on the left. The probability density function (PDF) per cluster is shown with respect to the indicated external information access measure given on the x -axis. Note that the cluster counts for the Job Rank column are the same as for the PhD Rank column; the annotations are omitted in the figure to avoid obscuring the distribution curves.

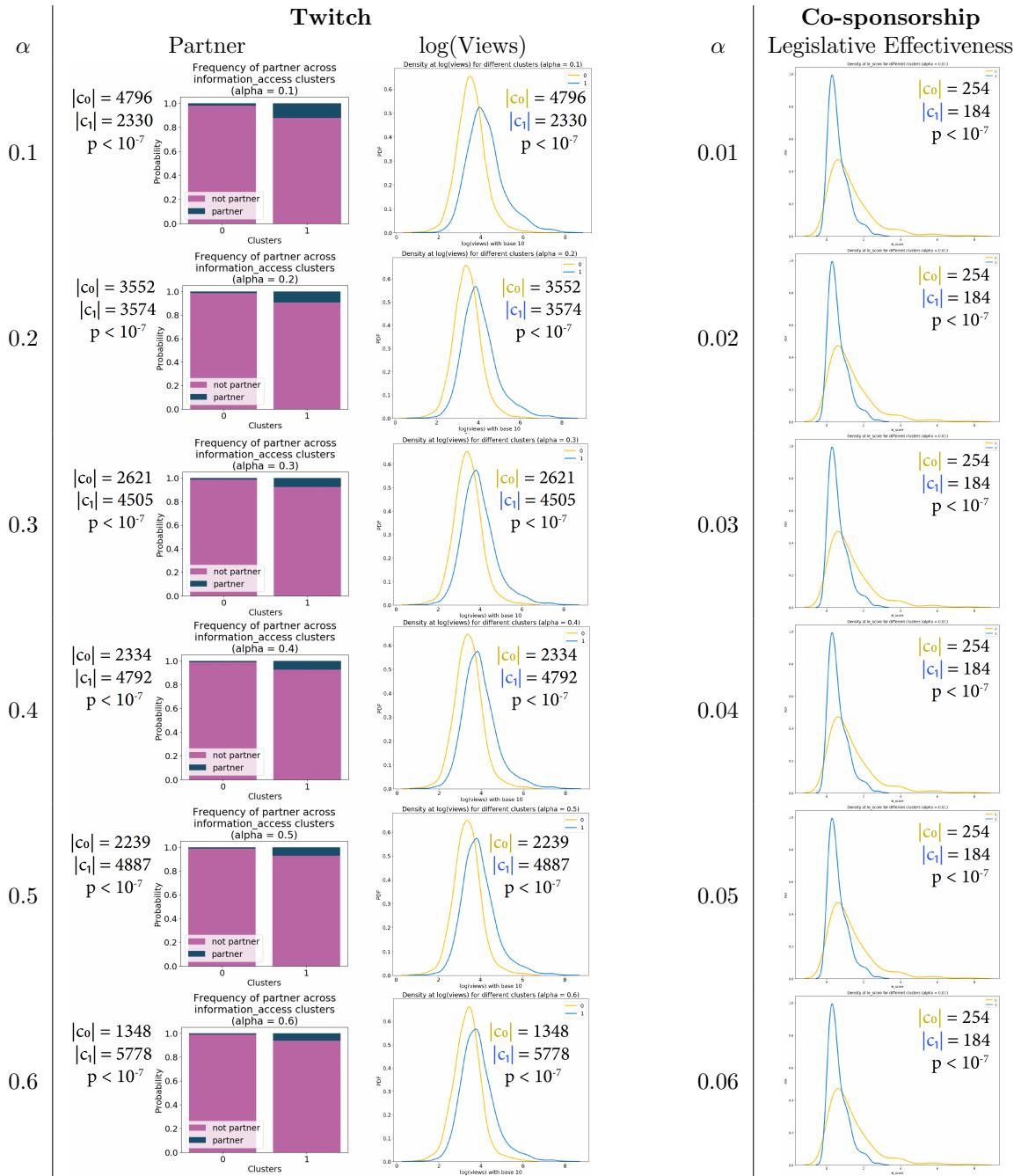


Figure 11: Results above are shown for the Twitch and Co-sponsorship datasets. Clusterings are given for α values as shown on the left. The probability density function (PDF) or composition per cluster is shown with respect to the indicated external information access measure.

C Comparison between different clustering methods

	α	DBLP	α	Twitch	α	Co-spons.
Spectral Clustering	0.1	~ 0	0.1	~ 0	0.01	~ 0
	0.2	~ 0	0.2	~ 0	0.02	~ 0
	0.3	~ 0	0.3	~ 0	0.03	~ 0
	0.4	~ 0	0.4	~ 0	0.04	~ 0
	0.5	~ 0	0.5	~ 0	0.05	~ 0
	0.6	0.02	0.6	~ 0	0.06	~ 0
	0.7	0.02	0.7	~ 0	0.07	~ 0
	α	DBLP	α	Twitch	α	Co-spons.
Fluid Communities	0.1	~ 0	0.1	~ 0	0.01	0.65
	0.2	~ 0	0.2	0.01	0.02	0.50
	0.3	~ 0	0.3	~ 0	0.03	0.28
	0.4	~ 0	0.4	~ 0	0.04	0.17
	0.5	~ 0	0.5	~ 0	0.05	0.07
	0.6	~ 0	0.6	~ 0	0.06	0.02
	0.7	~ 0	0.7	~ 0	0.07	~ 0
	α	DBLP	α	Twitch	α	Co-spons.
Louvain	0.1	~ 0	0.1	~ 0	0.01	0.85
	0.2	~ 0	0.2	~ 0	0.02	0.59
	0.3	~ 0	0.3	~ 0	0.03	0.27
	0.4	~ 0	0.4	~ 0	0.04	0.12
	0.5	0.01	0.5	~ 0	0.05	0.02
	0.6	0.03	0.6	~ 0	0.06	~ 0
	0.7	0.01	0.7	~ 0	0.07	~ 0
	α	DBLP	α	Twitch	α	Co-spons.
Role2Vec	0.1	~ 0	0.1	0.11	0.01	0.77
	0.2	0.37	0.2	0.15	0.02	0.55
	0.3	0.49	0.3	0.13	0.03	0.27
	0.4	0.44	0.4	0.12	0.04	0.13
	0.5	0.38	0.5	0.11	0.05	0.03
	0.6	0.22	0.6	0.05	0.06	~ 0
	0.7	0.20	0.7	0.04	0.07	~ 0
	α	DBLP	α	Twitch	α	Co-spons.
Core-Periphery	0.1	0.73	0.1	0.54	0.01	0.12
	0.2	0.19	0.2	0.16	0.02	0.11
	0.3	~ 0	0.3	~ 0	0.03	~ 0
	0.4	~ 0	0.4	~ 0	0.04	~ 0
	0.5	~ 0	0.5	~ 0	0.05	~ 0
	0.6	~ 0	0.6	~ 0	0.06	~ 0
	0.7	~ 0	0.7	~ 0	0.07	~ 0

Table 6: For each dataset, α -parameterized clustering, and k clusters, the above table gives the adjusted rand index indicating the difference between the resulting information access and the indicated clustering method. An adjusted rand index of 0 indicates only random agreement between clusterings while a value of 1 indicates an exact match. For legibility, adjusted rand index values less than 0.01 are shown as ~ 0 . Fluid communities method given values are the mean across 10 random runs.

D Large networks

		Random sampling													
		Number of sampled seeds													
α		5	10	15	20	25	30	35	40	45	50	55	60	65	70
0.1	0.00	0.81	0.88	0.89	0.87	0.83	0.87	0.92	0.93	0.92	0.92	0.88	0.93	0.93	0.92
0.2	0.97	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.98	0.99	1	0.99	0.99	0.99	0.99
0.3	0.97	0.99	0.99	0.99	0.99	0.99	0.99	1	1	0.99	1	0.99	0.99	0.99	0.99
0.4	0.99	0.99	1	0.99	1	1	1	1	0.99	1	1	1	1	1	1
0.5	1	1	1	0.99	1	1	1	1	1	1	1	1	1	1	1
0.6	0.99	1	1	1	1	1	1	1	1	0.99	0.99	1	0.99	0.99	0.99
0.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		Selection in order of PageRank													
		Number of sampled seeds													
α		5	10	15	20	25	30	35	40	45	50	55	60	65	70
0.1	0.89	0.90	0.93	0.92	0.93	0.92	0.93	0.95	0.94	0.94	0.95	0.95	0.95	0.95	0.94
0.2	0.99	0.99	1	1	1	1	1	1	0.99	1	1	1	1	1	1
0.3	0.99	1	1	0.99	1	0.99	1	1	1	1	1	1	0.99	1	1
0.4	1	1	1	1	1	1	1	0.99	1	0.99	1	0.99	0.99	0.99	1
0.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.6	0.99	0.99	1	0.99	1	1	0.99	1	1	0.99	0.99	1	1	1	1
0.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		Selection in order of degree centrality													
		Number of sampled seeds													
α		5	10	15	20	25	30	35	40	45	50	55	60	65	70
0.1	0.901	0.91	0.92	0.92	0.95	0.96	0.94	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95
0.2	1	0.99	0.99	1	0.99	1	1	1	1	1	1	1	0.99	1	1
0.3	1	0.99	1	0.99	1	1	0.99	1	1	1	1	1	0.99	1	1
0.4	1	1	1	1	1	0.99	1	1	1	1	1	1	1	1	1
0.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.6	1	1	1	0.99	1	1	0.99	0.99	0.99	0.99	1	1	1	1	0.99
0.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		Selection in order of betweenness centrality													
		Number of sampled seeds													
α		5	10	15	20	25	30	35	40	45	50	55	60	65	70
0.1	0.91	0.91	0.91	0.91	0.91	0.92	0.94	0.92	0.93	0.92	0.92	0.93	0.92	0.92	0.92
0.2	0.99	0.99	1	1	1	1	1	0.99	1	1	1	1	1	1	1
0.3	0.99	0.99	1	0.99	0.99	1	1	1	1	1	1	1	1	1	1
0.4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0.6	1	1	0.99	1	0.99	1	1	0.99	0.99	0.99	1	0.99	1	0.99	0.99
0.7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 7: Adjusted rand index values for the comparison between the information access clustering on the full representation for the DBLP network versus the information access clustering based on the shown selected number of seeds and seed selection strategies. Adjusted rand index values are 1 when two clusterings agree on all pairs of points.

D.1 Google Scholar dataset experimental setup details

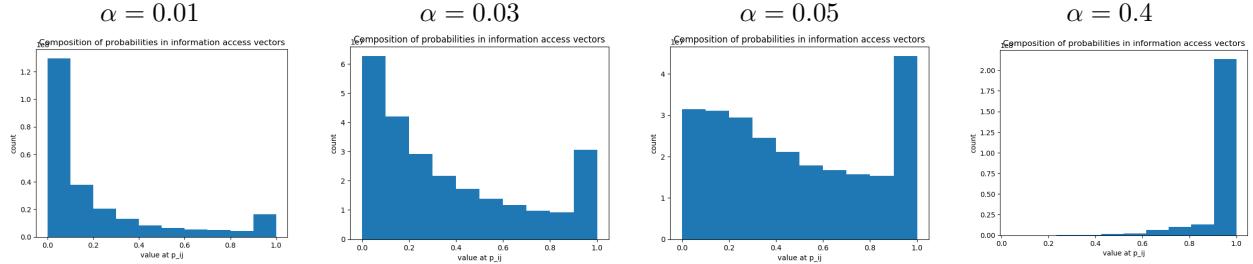


Figure 12: Histograms showing the prevalence of p_{ij} values within the information access signatures for the Google Scholar data based on α value.

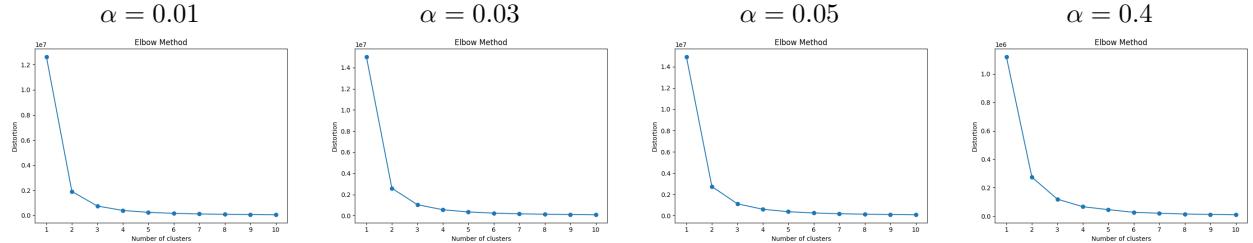


Figure 13: Elbow method plots for information access clustering on the Google Scholar dataset.

		Google Scholar									
		k									
α		2	3	4	5	6	7	8	9	10	
0.01		0.7448	0.6919	0.6630	0.6445	0.6317	0.6222	0.6154	0.6088	0.6040	
0.03		0.7154	0.6803	0.6613	0.6489	0.6403	0.6332	0.6264	0.6208	0.6171	
0.05		0.7191	0.6918	0.6782	0.6702	0.6629	0.6590	0.6547	0.6505	0.6478	
0.40		0.9132	0.8993	0.8903	0.8910	0.8851	0.8785	0.8786	0.8782	0.8717	

Figure 14: Silhouette values for varying α and k for the Google Scholar dataset for information access clustering. The largest silhouette value for each α is shown in bold.