doi:xx.xxxx/xxxxx

Linking Datasets on Organizations Using Half A Billion Open Collaborated Records

CONNOR JERZAK & BRIAN LIBGOBER Harvard & UCSD

Scholars of organizations often face challenges connecting datasets. Disparate sources rarely share identifiers or covariates. Therefore, researchers usually resort to exact or fuzzy string matching between different lists of organization names. Nevertheless, these techniques may struggle to find correct pairs. Widely used names for the same entity often have few characters in common (e.g., 'Chase Bank' and 'JPM'). In this letter, we offer an alternative. We build an organizational alias directory from over half-a-billion human-contributed records on LinkedIn. We do so by transforming these records into an alias network and applying an unsupervised Markov clustering algorithm for extracting high probability matches between organization names. The resulting directory contains publicly traded firms, NGOs, small businesses, and government agencies from across the world. We highlight the directory's value through an application on lobbying by publicly traded firms. We make our software available in an open-source R package ('LinkIt').

Word Count: 3,301

RECORD-LINKAGE AND ORGANIZATIONAL SOCIAL SCIENCE

As large datasets on individual political behavior have become more prevalent, scholars have focused increasing attention on the methodological problem of linking records from different sources (Enamorado, Fifield, and Imai 2019; Herzog, Scheuren, and Winkler 2010; Larsen and Rubin 2001). This task is important because, through it, researchers can obtain outcome or covariate data about survey respondents, campaign contributors, or voters that would have been costly or impossible to obtain in previous eras (e.g. Ansolabehere and Hersh 2012; Figlio et al. 2014; Bolsen, Ferraro, and Miranda 2014; Hill and Huber 2017). When unique identifiers are unavailable for linking datasets, this recent research has developed sophisticated record linkage algorithms which can find, with high probability, the same individuals in two datasets using stable characteristics such as birth year and race.

These developments have had less of an impact on scholarship concerning organizational entities such as corporations, universities, trade associations, think tanks, religious groups, non-profits, and international organizations. As with research on individuals, scholars of organizations also combine multiple data streams to develop evidence-based models. However, in addition to lacking shared unique numeric identifiers, such datasets also often lack common covariate data that form the basis for probabilistic linkage algorithms. Therefore, scholars rely heavily on exact or fuzzy string matching to link records.

To take a recent example from APSR, Crosson, Furnas, and Lorenz (2020) compare the ideology scores of organizations that have political action committees with ones that do not. The ideology scores are calculated from a dataset of interest group position taking made by a non-profit organization (Maplight), while the list of organizations with political action committees ultiamtely derives from Federal Election Commission records. Maplight and the Federal Election Commission may not refer to organizations in the same way and there is no covariate data that one can use to help with linkage. Matching records in this situation is “challenging” (p. 32), and the authors consider both exact and fuzzy matching. Ultimately, they adopt a combination of pre-processing and exact matching because of concerns about false positives, while acknowledging that they inevitably do not link all related records as a result. Indeed, the authors supplement the 545 algorithmic matches with 243 hand-matches, implying that their first algorithmic effort missed at least one in three correct matches.

The challenge that Crosson, Furnas, and Lorenz (2020) face is typical for scholars studying organization. Informally, our impression is that they have surmounted it more effectively, and perhaps with greater transparency, than typical. The example illustrates how, although powerful, string matching has fundamental limitations. While it can link records whose identifiers contain minor differences, it has trouble handling the diversity of monikers an organization may have. For example, “JPM” and “Chase Bank” refer to the same organization, yet these strings share no characters. Indeed, even human coders may struggle to connect records between some organizations. For example, string matching and research assistants both may fail to detect a relationship between Boalt Hall and Berkeley Law School, although the former was a common name for the latter until 2020. Such difficulties connecting data sources routinely hinder research on organizations, especially when attempting to link datasets in different source languages.

In this letter, we develop a tool that can help. First, we use the full LinkedIn database to construct a weighted graph indicating the probability that two organizational names share the same reference point. For example, "JPM" is one node in this graph, while "JP Morgan" is another. By counting the frequency with which each name is associated with shared organizational URLs (i.e. `https://www.linkedin.com/company/jpmorgan`), we generate the probability of connection between nodes. With these probabilities in hand, we then apply a community detection algorithm to this directed graph. This algorithm allows us to create a large-scale directory which can help researchers in linking datasets on both public and private organizations from over 100 countries and in dozens of languages. Intuitively, the directory uses the combined wisdom of millions of human beings with first-hand knowledge of these organizations. Often, it can succeed in connecting aliases where string and manual record-linkage approaches would struggle.

We illustrate the usefulness of this directory through an application involving lobbying data. An open-source package ("LinkIt") implementing computationally efficient matching using the directory has been posted online at `url-hidden`.

DATA SOURCE DESCRIPTION

User-contributed records from LinkedIn are our primary data source. We acquired these records from a vendor, Datahut.co, which stated that they were a complete site scrape circa 2017. The Ninth Circuit Court of Appeals established in HIQ Labs, Inc., v. LinkedIn Corporations (2017) the right of firms to obtain and market such data. The dataset contains about 350 million mostly unique profiles drawn from over 200 countries—a size and coverage consistent with LinkedIn's own estimates.

TABLE 1. Illustration of source data.

Full Name	Title	Organization	Profile Url	Start Date	End Date
Carole Baskin	CEO	Big Cat Rescue	http://www.linkedin.com/company/big-cat-rescue	1992-11-01	2017-04-16
Carole Baskin	Founder	Big Cat Rescue	http://www.linkedin.com/company/big-cat-rescue	1992-11-01	2017-04-16
Carole Baskin	Owner	Guardian Angel Land Trust		1981-01-01	2017-04-16
Doc Antle	Director	R.S.F.	http://www.linkedin.com/company/r-s-f/	1982-01-01	2017-03-14
Doc Antle	DIRECTOR	T.I.G.E.R.S.		1982-01-01	2017-03-14
Mario Tabraue	Co-Founder and President	Zoological Wildlife Foundation	http://www.linkedin.com/company/zoological-wildlife-foundation		2016-08-28

To construct the directory, we focus on the professional experiences posted by users. In each profile on LinkedIn, a user may list the name of their employer as free-response text. Users also may link this experience to the profile of their employer. They may also decline to do so or may make mistakes. Large, internally differentiated organizations often have multiple valid profile URLs. For example, ICPSR and the University of Michigan both have distinct organizational profile URLs on LinkedIn. Table 1 provides an example extract of the professional experience table for several public figures.

GRAPH CONSTRUCTION

User profiles on LinkedIn contain hundreds of millions of associations between organization names and URLs. We apply a fundamental law of probability and an independence assumption to average across the noise present in the name-to-URL data. We thereby obtain the most likely name-to-organization mappings.

To formalize our approach, we denote organization names by $a_1, a_2, \ldots a_N$ and consider $\Pr\left(a_i \mid a_j\right)$. This expression represents the probability that, given name a_j, the organization with name a_i is intended. We can re-write this probability as

$$\Pr(a_i \mid a_j) = \sum_{u \in \mathcal{U}} \Pr\left(a_i \mid u, a_j\right) \Pr(u \mid a_j), \tag{1}$$

where $\mathcal{U}$ is the set of all URLs and u is a specific URL from this set. It is straightforward to calculate $\Pr(u \mid a_j)$ as the proportion of times that URL u is selected by users entering organization name a_j. $\Pr\left(a_i \mid u, a_j\right)$ is tougher, but tractable if we make an independence assumption. We suppose that given a URL, the alias a_j adds no information about the probability that alias a_i is intended. This assumption is comparable to that behind the naive Bayes classifier, a common algorithm which often gives good predictive performance. Under this assumption, Equation 1 can be simplified:

$$\Pr(a_i \mid a_j) = \sum_{u \in \mathcal{U}_j} \Pr\left(a_i \mid u\right) \Pr(u \mid a_j), \tag{2}$$

where the quantity, $\Pr\left(a_i \mid u\right)$, can be calculated as the proportion of cases where a user selecting organization URL u has written organization name a_i.

Using this approach, we now can calculate a probability matrix, $\mathbf{P}$, where entry i, j is $\Pr\left(a_i \mid a_j\right)$. This matrix induces a directed graph on organization names as nodes. The matrix P is sparse since most organization names are never intended to refer to the each other.

To illustrate the calculation of $\Pr\left(a_i \mid u\right)$, suppose we wished to know the (hopefully low) probability that "Michigan State University" is an alias for "University of Michigan." The term "University of Michigan" occurs in 63,196 employment experiences in our source data. It is associated with some 47 organization URLs. The URL, `https://www.linkedin.com/company/university-of-michigan`, abbreviated here as URL-UM, covers more than 99.5% of these experiences (to keep the illustration simple, we ignore the impact of these other 0.05% of URLs). URL-UM appears 75,462 times in the raw data. There is one instance where someone wrote their employer was "Michigan State University" and (presumably incorrectly) selected this URL. Thus, the probability of "Michigan State University" given "University of Michigan" is about 1 in 75,000. By contrast, the organization name "ICPSR" appears with the URL-UM eight times, so the probability of "ICSPR" given "University of Michigan" is far greater than the probability of "Michigan State." This example shows how our proposed approach combines information from many people in order to reduce the impact of out-of-concensus users.

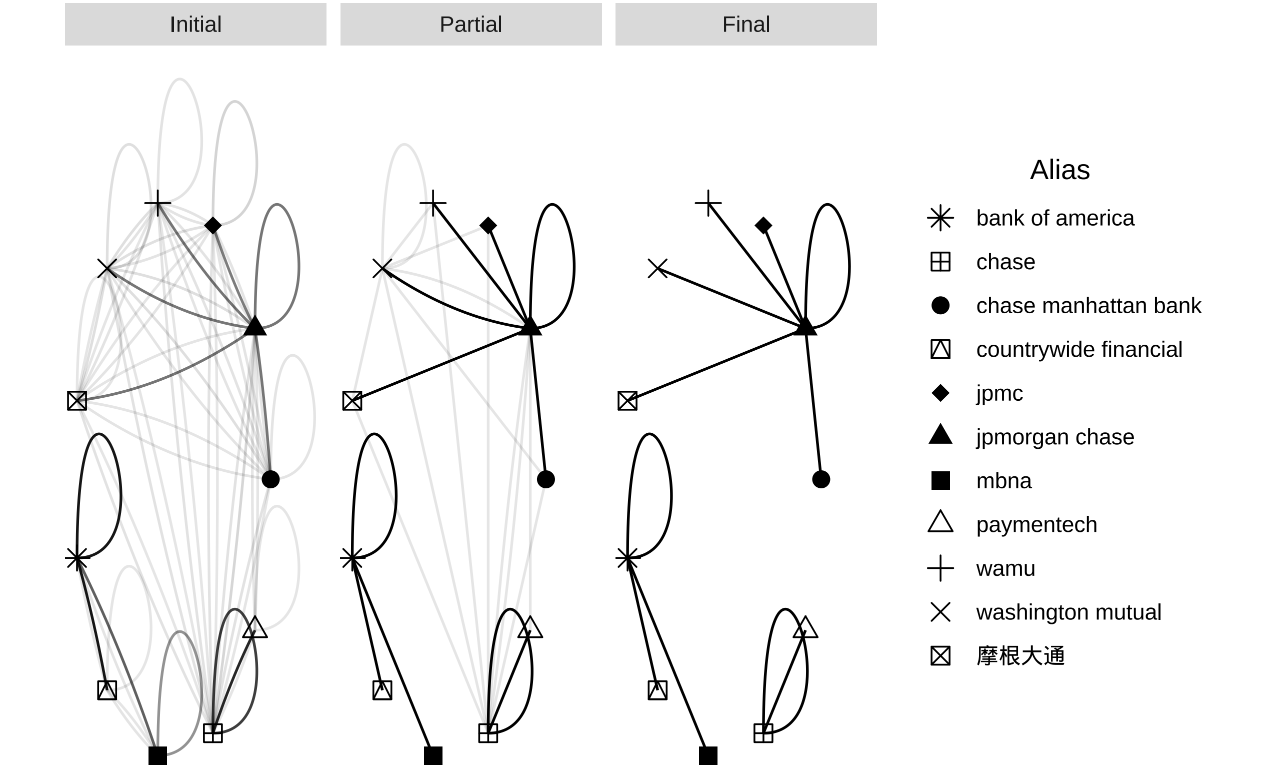
COMMUNITY DETECTION

We now seek “communities” or clusters of densely connected names in the graph induced by the probability matrix, $\mathbf{P}$. These clusters of connected names are what our LinkIt software uses to merge datasets on organizations when those sources use disparate names for the same entity. We consider computationally efficient algorithms given the size and sparsity of this probability matrix. Moreover, since using network models for organizational linkage is novel, we also focus on algorithms that are simple and transparent. The Markov clustering algorithm—first proposed by Van Dongen (Van Dongen 2008) and popular in the bio-informatics context—satisfies our criteria. It is worth noting that there is a vast literature on community detection in networks that offers alternative methods which can be explored in future research (Rohe, Chatterjee, and Yu 2011).

The intuition for Markov clustering arises from an observation about finite state Markov processes, which are used in modeling event sequences where the system evolution at one time depends only on its state at the previous point. Suppose a traveler exploring a network starts out at node i. If the i, j-th entry of matrix $\mathbf{P}$ gives the probability of going to node j from node i, then $\mathbf{P}_n = \mathbf{P} \times \mathbf{P} \times \ldots \times \mathbf{P}$ defines the probability of being at node j after n transitions. If n is small, the probability the traveler has remained in i’s community is high. As n increases, however, the probabiliy of exiting i’s community grows. Once one moves from one densely connected cluster to another, the probability of returning to the first cluster quickly drops. In the limit as $n \to \infty$, the “clustered” aspect of short-term transition probabilities is lost. Markov clustering tries to prevent the loss of information about neighborhoods through a remarkably simple calculation. After starting off with an initial $\mathbf{P}$ matrix and visiting a few nodes (done by taking $\mathbf{P} \times \mathbf{P} \times \cdots \times \mathbf{P}$), we simply raise all the terms to some power, renormalize every row to form $\tilde{\mathbf{P}}$, and then iterate again (by taking $\tilde{\mathbf{P}} \times \tilde{\mathbf{P}} \times \cdots \times \tilde{\mathbf{P}}$). This process rapidly converges upon stable clusters (Van Dongen 2008), where the number of clusters is determined from the data and not by researchers.

Figure 1 illustrates the process on a subset of our data. Darker shades reflect heavier weights in $\mathbf{P}$. Some links are much stronger than others. In the initial weighting, two cliques reflect

FIGURE 1. **Illustration of Markov clustering.**



a set of names associated with "jp morgan chase." Another reflects names associated with "bank of america." However, these initial links are dense, making it difficult to distinguish one cluster of aliases from another. As the algorithm iterates, some links weaken and disappear while others strengthen. Eventually, each node links to exactly one other node. Notably, the final cliques contain lexographically dissimilar nodes that do indeed belong in the same cluster. For example, the "chase" clique contains "wamu", "paymenttech", and "摩根大通" which are all chase affiliates. The "bank of america" clique includes "countrywide financial" and "mbna," both under the Bank of America umbrella. By examining the consequential behavior of LinkedIn users, this method allows us to find in a data-driven way the most likely matches between the organizational aliases found in political science datasets.

VALIDATION AND APPLICATION

We illustrate the use of the organizational directory on a record linkage task involving lobbying and the stock market. Libgober (2020) shows that firms that meet with regulators tend to receive positive returns in the stock market after the regulator announces the policies on which they lobbied. These returns are significantly higher than the positive returns experienced by market competitors and firms that send regulators written correspondence. Matching of meeting logs to stock market tickers is burdensome because there are almost 700 distinct organization names described in the logs and around 7000 public companies. Manual matching typically involves research on these 700 entities using tools such as Google Finance. While the burden of researching seven hundred organizations in this fashion is not enormous, Libgober (2020) only considers meetings with one regulator. If one were to increase the scope to cover more agencies, or lobbying in Congress, the burden could become insurmountable.

We show how the incorporation of our directory into the matching process can improve performance, treating the human coded matches in Libgober (2020) as ground truth. Formally, fuzzy matching is a procedure that calculates string dissimilarities between all pairs of names in two datasets. Two entries are declared a match if the dissimilarity is below an acceptable threshold. The procedure used in the LinkIt software developed here has two steps, the first of which is the same as simple fuzzy matching. In the second step, we perform fuzzy matching of each dataset to the directory constructed from LinkedIn data. The directory provides a canonical numeric identifier for each successful match in stage two. This shared numeric identifier provides the basis for a second round of matching where hard-to-identify matches have a better chance to be found.

For any particular acceptance threshold, the LinkIt procedure finds at least as many matches as the simple procedure. It is not always the case, however, that the additional matches from the LinkIt step represent a performance gain. Some of these matches might be true positives (improving performance), while others might be false positives (hurting performance). Moreover, each procedure may be applied using a different acceptance threshold. As we show, a significant advantage of the supplemented procedure is that it can achieve a similar number of positive matches as fuzzy matching with a lower acceptance threshold. Lower acceptance thresholds result in fewer false matches, so are generally preferred.

In what follows, we focus on string matching using the cosine distance measure (the Jaccard distance measure produces similar results). For two strings a and b, the cosine measure is constructed as follows. Let A and B reflect the decomposition of a and b respectively into a binary vector whose entries are 1 if a given q character combination (known as q-gram) is present and 0 otherwise. Then

$$d(a,b) = 1 - \frac{\sum_{d=1}^{D} A_i B_i}{\sqrt{\sum_{d=1}^{D} A_i^2}\sqrt{\sum_{d=1}^{D} B_i^2}},$$

where D is the total number of q-grams present in A or B. If all q-grams co-occur within A and B, the measure is 0. If none co-occur, the measure is 1. Following (Navarro and Salmela 2009), we set $q = 2$. The results are not sensitive to this choice.

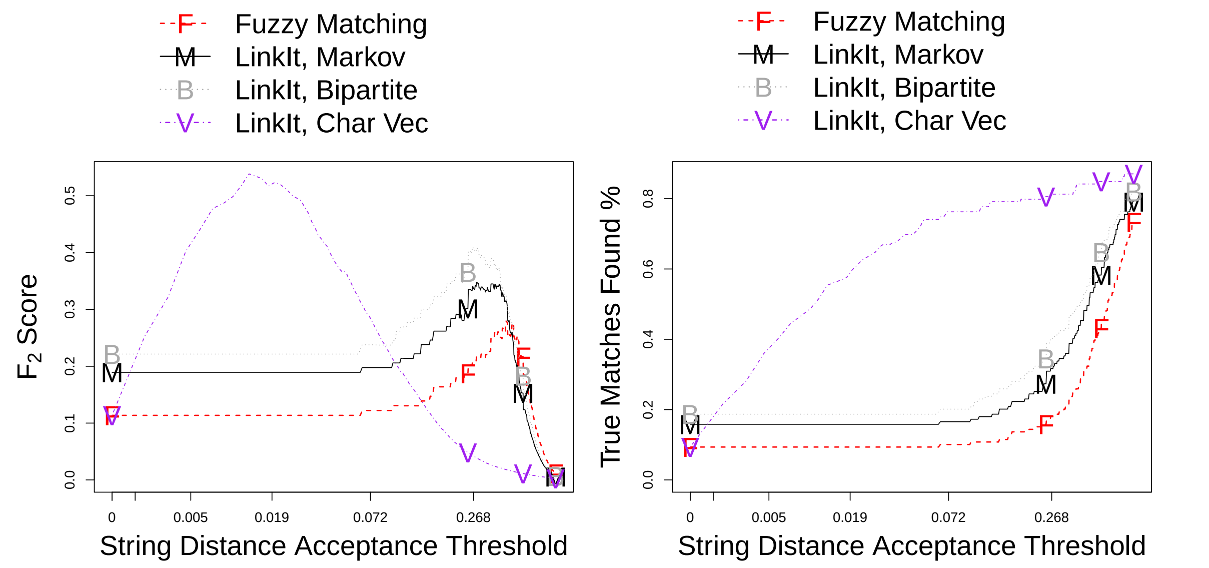
We examine two measures of performance—(a.) the number of true matches found and (b.) a measure which considers the presence of true positives, false positives, and false positives which is known as the F_β score. This score is defined formally as

$$F_\beta = \frac{(1+\beta^2)\cdot \text{true positive}}{(1+\beta^2)\cdot \text{true positive} + \beta^2 \cdot \text{false negative} + \text{false positive}} \tag{3}$$

In the best case scenario, the F_β score is 1, which occurs when all true matches are found, with no false negatives or false positives. In the worst case scenario, the score of 0, which occurs when no true positives are found. If an algorithm obtains some positives matches but many more false negatives or false positives, the measure also approaches 0. The parameter β controls the relative costs of false negatives as compared with false positives. If β>1, false negatives are regarded as less costly than false positives; if β<1 then the reverse. In the matching context, errors of inclusion are typically less costly than errors of exclusion because the list of succesful matches is usually shorter and easier to double-check than the list of non-matched pairs. Therefore, the particular version of F_β we focus on is F_2.

Figure 2 shows that the directory-based approach yields higher F_2 score across the acceptance threshold range, and also returns more true positives. The percentage gain in performance is highest when the fuzzy acceptance threshold is near 0, which is often preferable in practice as higher acceptance threshold can yield low quality matches. This result illustrates how our

FIGURE 2. We find that dataset linkage using the Markov clusters on the LinkedIn alias network obtains superior performance both when adjusting for the rate of false positives (left panel) and also when counting only the raw percentage of correct matches obtained (right panel).



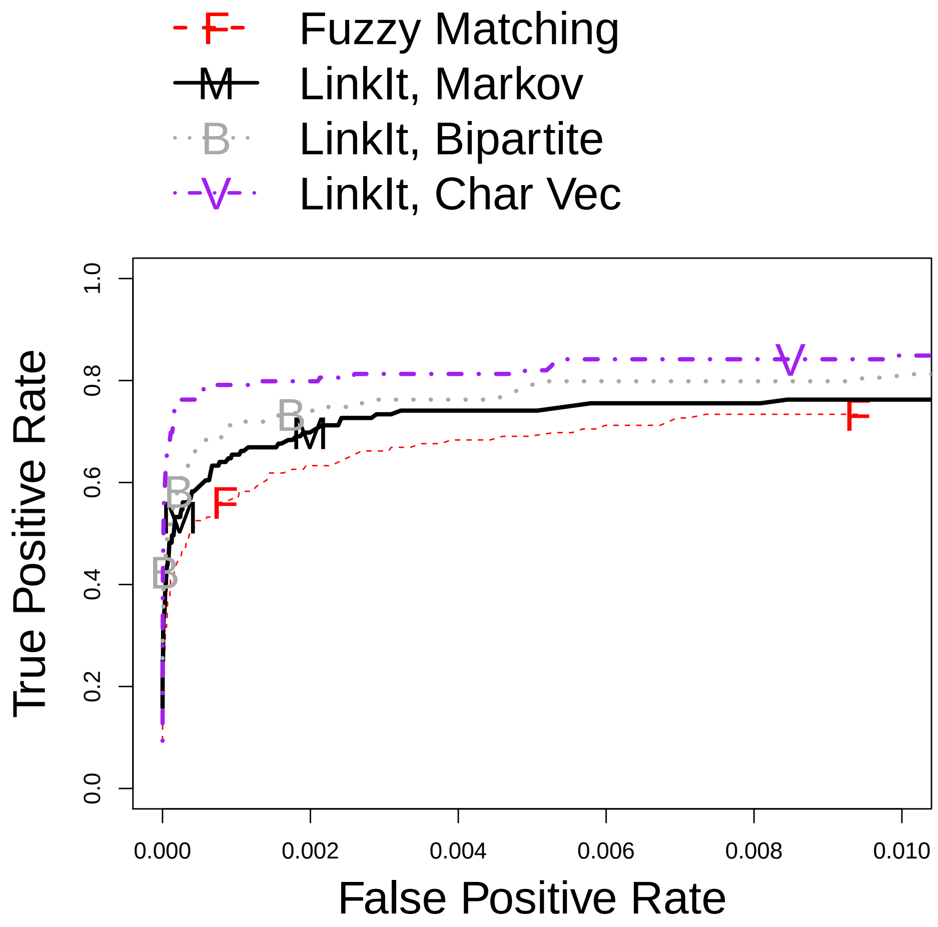
directory-based method yields superior performance compared to a leading alternative method for obtaining organizational matches in the common use case when researchers do not have access to shared covariates across datasets.

It is also instructive to consider an example from this linkage task where fuzzy matching failed but the LinkIt approach has success. In particular, fuzzy matching fails to link the organizational log entry associated with “HSBC Holdings PLC” to the stock market data associated with “HSBC.” Their string distance using the cosine measure is 0.57, which is much higher than the distance of “HSBC Holdings PLC” to its fuzzy match (0.13 for “AMC Entertainment Holdings, Inc.”). “HSBC Holdings PLC”, however, has an exact match in the LinkedIn-based directory, so that the two organizations are successfully paired even with a fuzzy-matching threshold of 0 according to the LinkIt procedure.

DISCUSSION

Organizational data often lacks common covariate data (such as race or gender), making it difficult to apply probabilistic linkage methods and motivating the widespread use of fuzzy

FIGURE 3. ROC description.



matching algorithms. We have presented a new tool for improving the matching of organizational entities using over a half-billion open collaborated employment records from a prominent online social network. This approach can match organizations which contain no common linguistic identifiers or which are written in different languages. We validate the approach on an example task linking organizational meeting logs to stock market tickers. We show superior performance to the most common alternative method (fuzzy matching). As a novel application of community detection to this area, we have erred on the side of simplicity and transparency in extracting an organizational directory from these open collaborated records. Future work may attempt more sophisticated approaches to graph construction, community detection, and matching calibration using this unique data source.

REFERENCES

Ansolabehere, Stephen, and Eitan Hersh. 2012. “Validation: What big data reveal about survey misreporting and the real electorate”. Political Analysis 20 (4): 437–459.

Bolsen, Toby, Paul J. Ferraro, and Juan Jose Miranda. 2014. “Are voters more likely to contribute to other public goods? Evidence from a large-scale randomized policy experiment”. American Journal of Political Science 58 (1): 17–30.

Crosson, Jesse M, Alexander C Furnas, and Geoffrey M Lorenz. 2020. “Polarized Pluralism Organizational Preferences and Biases in the American Pressure System”. American Political Science Review. accepted.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2019. “Using a probabilistic model to assist merging of large-scale administrative records”. American Political Science Review 113 (2): 353–371.

Figlio, David, et al. 2014. “The effects of poor neonatal health on children’s cognitive development?” American Economic Review 104 (12): 4205–4230.

Herzog, Thomas H., Fritz Scheuren, and William E. Winkler. 2010. “Record linkage”. Wiley Interdisciplinary Reviews: Computational Statistics 2 (5): 535–543.

Hill, Seth J., and Gregory A. Huber. 2017. “Representativeness and Motivations of the Contemporary Donorate: Results from Merged Survey and Administrative Records”. Political Behavior 39 (1): 3–29.

Larsen, Michael D., and Donald B. Rubin. 2001. “Iterative automated record linkage using mixture models”. Journal of the American Statistical Association 96 (453): 32–41.

Libgober, Brian. 2020. “Meetings, Comments, and the Distributive Politics of Rulemaking”. Quarterly Journal of Political Science.

Navarro, Gonzalo, and Leena Salmela. 2009. “Indexing variable length substrings for exact and approximate matching”. In International Symposium on String Processing and Information Retrieval, 214–221.

Rohe, Karl, Sourav Chatterjee, and Bin Yu. 2011. “Spectral Clustering and the High-Dimensional Stochastic Blockmodel”. The Annals of Statistics 39 (4): 1878–1915.

Van Dongen, Stijn. 2008. “Graph clustering via a discrete uncoupling process”. SIAM Journal on Matrix Analysis and Applications 30 (1): 121–141.