Collapsing Corporate Confusion

Leveraging Network Structures for Effective Entity Resolution in Relational Corporate Data

Tim Marple
Department of Political Science
University of California, Berkeley

Bruce Desmarais
Department of Political Science
Pennsylvania State University

Kevin L Young
Department of Political Science
University of Massachusetts, Amherst

Abstract*— In this paper, we introduce a novel battery of classifiers to resolve artificial inconsistencies among entity names within large datasets. Using data on the corporate sector, we describe the logic underlying a relational approach to entity resolution, and its importance for data acquisition, feature extraction, and integration. We subsequently leverage the relational structure of BoardEx employment data to assess the efficacy of these methods as compared to a ground-truth sample of coded name inconsistencies. We show that these methods hold significant promise for cleaning artificial distinctions in entity names via enrichment from integration with external data, and further demonstrate the effect of such resolution on the accuracy of extracted network topology features. We conclude with implications for existing findings and steps for future work.

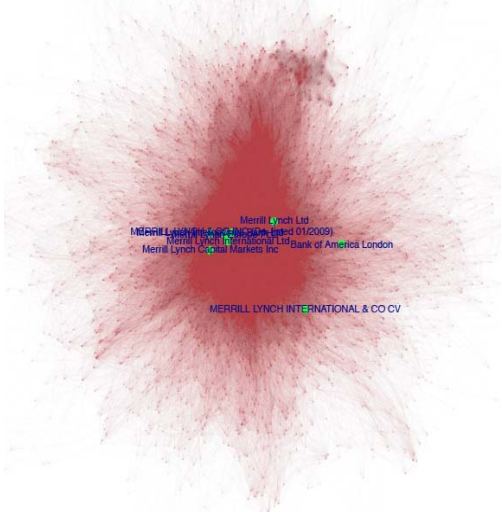
Keywords-- entity resolution; network methods; corporate data; BoardEx; rare events regression

I. INTRODUCTION

The quality of big data is significantly undermined by errors in recorded entity names, yet this issue pervades a great number of existing large datasets. Inconsistencies in entity names impede on quality data analysis along three crucial dimensions: the acquisition of accurate data, the extraction of relevant features, and the integration of big data. As such, a significant issue facing researchers using big data is the ability to resolve errors in their datasets which do not reflect what they aim to describe. Here, we present a novel approach to such entity resolution in large datasets premised on leveraging the relational structure of network data to collapse artificially distinct entities. We deal with a particularly challenging case of entity resolution - corporate network data. Because of the prominence of the modern corporation in the organization of wealth and power in the contemporary world, working with corporate network data is useful for approaching many topics in economics, sociology, political science and public policy. Yet corporate organization is not characterized by simple, singular entities but rather a multiple array of organizations linked in ownership and subsidiary chains, involving shell companies and legally remote hubs for various activities.

Large proprietary databases of corporate network data are now available and present researchers with manifold opportunities for analysis. Yet the logic of modern corporate competition encourages a complex differentiation of corporate entity names, and nature of the corporate network data and of subsequent entity resolution problems and solutions must be understood in this context. In other words, entity ambiguity in corporate network data are not simply an artifact of bad or patchy data but rather reflect the nature of the corporate form itself. Because corporate activity spans many different jurisdictions, and thus different legal regulatory regimes, there is often a strong incentive to differentiate the firm to deal with these and other factors [4].

For example, Bank of America is composed of 229 different legal entities, including subsidiaries and shell companies, and not all of these contain permutations of the main corporate name in their official title. Figure 1 below shows an illustration of this problem. Figure 1 is a large employment network consisting of the second-degree employment ties for three important government institutions: the US Securities and Exchange Commission (SEC), the US Federal Reserve Board, and the Bank of England. It highlights in green nodes the firms that are part of the Bank of America corporate group. As this illustration points out, the location of 'Bank of America' in the network is ambiguous because of the multiple entities therein; many of which do not contain the components of the main name.



Network of Second-Degree Employment Ties, with Bank of America-linked Firms Highlighted in Green.

In this paper we describe novel relational entity resolution methods developed to address large corporate network dataset inconsistencies such as these. We utilize data from a large proprietary dataset used frequently by social scientists and business organizations [13] [15] [17], called Boardex. The entity resolution process we describe can be used in a variety of settings where corporate network data are concerned. We first describe existing approaches to entity resolution for large datasets, identifying two core models. Second, we describe our novel classifiers built to resolve entity ambiguity in our test dataset, BoardEx, demonstrating the utility of external datasets for accurate data acquisition. Third, we provide point estimates of precision, recall, and F1 scores for each classifier and their cumulative application, using a coded ground truth sample of entity distinctions. Fourth, we employ a rare-events logistic regression model to demonstrate the efficacy of the classifiers. Finally, we present effects of these classifiers on network topology measures to show the importance of entity resolution for accurate feature extraction. We conclude with implications for use of this type of big data in broader research projects.

II. Existing Approaches to Entity Resolution

Entity ambiguity is a common problem when analyzing databases in which character strings serve as ad-hoc keys, and no centralized key system is used when the data is input to the database. Some example database types for which methods for entity resolution have been developed include company sales data [12] [8], bibliometric databases [22] [7] [18], and law enforcement records databases [23] [9] [21]. These application domains all differ in terms of the overall structure and contents of data, but share the fact that the data is generated in a decentralized manner, and at a volume that makes it impossible to check each record by hand.

The methodological approaches to entity resolution fall into two general classes, which are sometimes used individually, but are typically combined. One class of methods involves various forms of character string matching, in which strings are combined if they are similar based on some string distance criterion [16] [1] [10]. A commonly applied metric is the edit or "Levenshtein" distance between two strings. The edit distance between two strings is given by the minimum number of characters that would need to be changed in one string to realize the other string [24]. A straightforward approach to entity resolution with edit distance is to classify entities as equivalent if they fall under a specified distance threshold [3] [10].

The second class of methods involves the analysis of the patterns of connections among the records in the database [14] [25]. Consider, for example, two authors with very similar names (e.g., Jane Smith, and Jane A. Smith). If they are both recorded as co-authoring with similarly-named co-authors, and/or authoring articles with very similar abstract contents, we have substantial evidence that Jane Smith and Jane A. Smith are different names for the same author entity. One example approach in this class is the use of network community detection [11] ---the discovery of groups of entities that are all closely connected to each other---to find entities that could potentially be equivalent [19] [25] [20].

III. Novel Relational Entity Resolution Methods

The entity resolution model presented here leverages the logic of ties between firms in a social network to collapse artificially distinct entities. These classifiers serve to effectively clean and enrich BoardEx data. In contrast with existing methods, which primarily treat names as the resolution mechanism, our approach uses relational data within social networks of nodes which may or may not require aggregation. The BoardEx data we use to test this approach includes data on many individuals' employment histories at a variety of firms across the world. As such, it allows for the construction of a two-mode network between individuals and organizations with edges in the network indicating an employment tie. This data structure further permits the construction of a one-mode organization to organization network in which ties constitute shared employees across organizations. Because this large network includes all unique organization names within the dataset as individual and distinct nodes, its relational structure offers a key by which artificial distinctions can be collapsed.

We argue that local communities of nodes are far likelier to include entity redundancies as a byproduct of the nature of corporate ties; firms which are recorded separately but which are actually the same should theoretically share disproportionately more employees than firms which are truly distinct. This can be the result of entity name changes over time, which are often well-recorded and serve as an excellent starting point for disambiguation. This can also be a result of employee transfers, which would retain an individual under the same theoretically relevant corporate structure, but record it as distinct. To resolve this issue, the classifiers proposed here involves six basic tasks to collapse artificially distinct nodes: collapsing firms based on recorded name changes, resolving corporate ownership hierarchies with external data, resolving node redundancy within network communities, resolving node redundancies within individual nodes' neighborhoods of shared network ties, and finally collapsing hand-coded, artificially distinct entities involved in the largest ties in the network.

Across the BoardEx dataset, firms were often recorded with their then-current names, and their former names in parentheticals based on when they were delisted and renamed. For example, TD Bank may have been recorded as: 'TD Bank (TD Banknorth prior to 09/2009)'. To account for this, we first cycled across every unique firm name and identified the 'prior to' or 'delisted' former name, if one existed. We subsequently cleaned every instance of firms with that former name in the dataset by replacing them with the newer name identified in the cell, omitting the parentheticals. As such, the example offered above would be reduced to 'TD Bank', and every instance of 'TD Banknorth' was converted to 'TD Bank', all without the data on former names and dates. Specifically, this comports with the logic of firm ownership as sufficient to aggregate otherwise distinct organizations, especially in the case of measuring network distances. It further serves as a basic, non-relational starting point to homogenize the entity data and prepare it for comparison against external datasets.

The second step in this process incorporated external data sources to clean and enrich the data, namely the corporate hierarchies of the firms within the BoardEx dataset. We used the corporate hierarchies of the 500 largest firms on a global ultimate owner basis, consisting of hierarchical ownership chains which allow for the identification of any firm within it and the replacement of its name with the highest firm on the chain. For example, 'TD Bank Mortgage Holdings' may be a subsidiary of 'TD Bank', but would be recorded in a social network as two distinct nodes. To collapse firms which fell within the same ownership chain, we iteratively cycled through each of the 500 largest global ultimate owner chains, and if a constituent firm name was found within the BoardEx data, it was replaced with the name of the firm at the top of that hierarchy. This process effectively reduced hierarchical incarnations of the same firm in our dataset via text matching for accurate integration with alternative, similar datasets. Following this step, we removed a broad population of corporate suffixes and acronyms as collected by two research assistants, in order to more cleanly match entity names within the remaining entity resolution classifiers[1].

The crux of this approach, and its next two steps, is its manipulation of social network logic to reduce the highest number of artificially distinct firm names. We leveraged two relational structures which appear within networks: communities and nodes' ego-networks. For both of these techniques, we first built a set of firm names' first words which uniquely identified a specific firm, and no other. These were gathered through a tabulation of firm name first words, each entry in which was hand-coded for its ability to uniquely identify a specific firm. Examples from this set include: Barclay's, KPMG, Toyota, and Wells Fargo. This list was checked by two coders, and only first words which both coders deemed to be unique identifiers of a given firm were retained within this replacement set.

The two network structures were manipulated in the same way. Each generated a set of unique nodes within the network; in the case of communities, these nodes were all members of the same computationally identified multilevel community [2], whereas ego-networks were the neighbors of each unique node in the network. In each group of entities, we first identified each member whose first word was within the approved list described above, and if two or more such nodes were present, their names were changed to the approved first word within the list. Then, if two or more nodes within each group shared the same first fifteen characters of their names, these fifteen characters were used to replace their names within the network. In this way, we leverage human knowledge of prominent firms and the relational structures of the social network to reduce artificial distinctions among firms and thereby clean the inconsistencies from the dataset.

The final step required troubleshooting the results of the first four. This required reviewing the network for any ties between intuitively identical firms which the classifiers described above missed. To accomplish this, we manually coded the connections in the network whose weights, or in this case the number of shared employees, was in the top 0.01% quantile. In total, this involved the manual coding of 1,829 edges, among which we identified 352 instances of ties between artificially distinct nodes. This allowed not only for the removal of those edges within the network, but also for the collapse of the names within those edges; the problematic edgelist served as a final key for firm aggregation. These edges were assessed by two human coders, and only the edges in the network which both coders deemed to be problematic were used. While this had the fewest firm aggregations of all prior steps in this method, it also solved some of the more pernicious problems.

[1] A dataset of common suffixes was collected from Wikipedia by two research assistants, and was cross-checked for meaningful content it may incidentally remove.

IV. Estimating Efficacy with Point Estimates

A. Ground Truth Sample

To gauge the efficacy of these classifiers, we employed a precision-recall test on their resolution for a hand-coded 'ground truth' sample from the BoardEx dataset. This first involved a random sample of 500 firms from the dataset, drawing from the >800,000 entities available. This then involved the hand-collection of those firms' alternate incarnations within the entire dataset by human coders searching across two corporate datasets, integrating external data to test the accuracy of our cleaning and enrichment processes. Using identical standards for adjudicating a match after in-person training, the coders searched both the BoardEx and LexisNexis Corporate Affiliations [6] datasets for any alternative version of a firm name within the ground truth firm sample. This typically involved 'lemmatizing' the entity name to identify other entities with the same basic root text in their names, and adjudicating with existing or external knowledge on its actual inclusion in or exclusion from the population of alternatives for each of the initial 500 samples firms. These alternative names were then searched within the BoardEx dataset to identify the unique firm identification code, the collection of which served as a population of accurate resolutions for each sampled firm. The mean number of alternate names found per firm was 20 with a median of 6; of the sample, 209 had no alternates.

B. Point Estimates of Precision, Recall, and F1 Scores

TABLE I. Precision, Recall, & F1 Scores

Precision, Recall, and F1 Results			
Entity Resolution Classifier	***Precision***	***Recall***	***F1***
Prior To/Delisted Removal	0.0087	0.0042	0.0057
GUO Corporate Hierarchies	>0.0000	>0.0000	>0.0000
Network: Communities	0.0116	0.0832	0.0204
Network: Shared Ties	0.0361	0.0429	0.0392
Problem Edge Removal	>0.0000	>0.0000	>0.0000
All (cumulatively applied)	0.0117	0.1187	0.0213

To initially assess these classifiers' efficacy in cleaning the data, point estimates for precision, recall, and their harmonic mean (F1 scores) were generated for the results of each method compared against the ground truth. This involved first constructing an entity-entity network for each method and the ground truth data, wherein a connection between entities indicates a dyad collapse. The adjacency matrices for these networks were then generated, omitting all columns without entities in the ground truth sample, to isolate their connections independently of those among firms which were not in the ground truth. Finally, because these matrices are so large (1,754,157 x 4,531 cells), precluding simple conversion to vectors, the population of non-0 row-column positions was collected for each matrix. Their comparisons to the ground truth row-column position population allowed for the point estimation of precision, recall, and F1 scores for each method. These scores are presented in Table I for the results of the classifiers.

In the table above, the final row indicates a data-informed, cumulative application of the classifiers described in this paper. These were applied in the order they are presented in the table above, such that the problem edge removal acts upon the shared ties results, which acted upon the communities results, and so forth. This sequence was chosen for its appropriateness for the BoardEx dataset and especially for the task at hand, namely the resolution of firms to their overarching corporate identity. As such, the values for this row indicate one of many possible combinations of these classifiers, but importantly demonstrate the collective efficacy of results compared to their individual, independent application to the dataset. Various research projects may find it fruitful to apply these classifiers, or any subset thereof, in varying sequences which may more appropriately suit their data or question. For example, it may be worthwhile to first search problematic edges and subsequently apply the relational classifiers, if one expects a different logic of entity duplicity in the observed corporate data.

These results indicate a relatively high rate of accurate name replacement by the entity resolution methods designed and applied here. Specifically, the relational network-based classifiers have by far the highest precision and recall, lending credence to the notion of leveraging ties for accurate data cleaning. Cumulatively it is clear there are interaction effects, visible in the diminished precision of the network-based approaches, though their cumulative recall is still high, capturing over 11% of true merges. While the values for GUO and problem edge classifiers are quite low, this may be reflective of the significantly lower likelihood of the firms to which they apply being present in a random sample of 500. These methods can be understood as independent approaches to resolving entity names; as such, to gauge their effects on producing the likelihood of a match, we next estimate a rare events regression model predicting these classifiers' successful identification of ground truth merges.

V. Estimating Efficacy with Rare Events Regression

The next step in estimating the efficacy of these classifiers was the construction of a rare events regression model to estimate their independent effects on correctly identifying a match. We first generated adjacency matrices for the ground truth sample and each method, with each cell indicating no match (0) or a match (1) suggested by the respective approach. Using the Zelig package in R [5], we used the predicted matches from each method to predict a swap in the ground truth sample. Because the computational burden of estimating the full adjacency matrix (~8 billion cells), we instead included all cells which had a match in the ground truth or any method, and subsequently imputed an additional 80 million zero-spots to resemble the population of non-merges against which these merges were to be compared. The population average of merges was therefore slightly less than 100 times inflated, though the tau corrector in the regression model accounted for this. The core model is presented below in Table III. Again, the network-based mechanisms significantly outperform most of their counterparts based in data cleaning and enrichment, as evidenced by their significance and larger coefficients. The removal of prior names, however, holds a comparable effect in identifying accurate dyad collapses.

TABLE II. Rare Events Logistic Regression Model

	Dependent variable:
	Ground Truth Entity Collapse
Prior-Delisted	4.8376***
	(0.2393)
GUO	-0.6595
	(36.991)
Communities	4.0293***
	(0.0530)
Shared Ties	2.7999***
	(0.0733)
Problem Edge	5.2815
	(36.968)
Constant	-13.7271***
	(0.0107)
Observations	80,000,000
Akaike Inf. Crit.	188,803
Note:	*p<0.1; **p<0.05; ***p<0.01

The communities approach holds the strongest predictive power for identifying correct firm matches of the two relational approaches, and the shared-ties approach is similarly significant with slightly lower classification efficacy. Interestingly, the GUO method has a negative coefficient, though the value is marginal and highly variable, suggesting instead a dearth of cases in which it was applied within the ground truth sample. The same is likely true for the problem edges classifier, though in this case it does have a large and positive coefficient, despite statistical insignificance. Broadly, the relational classifiers outperform their counterparts in both point estimates and in regression predictions. Each is significant well below the 0.001 level, and has a high log-odds of suggesting an accurate dyadic collapse between artificially distinct network nodes. This suggests that their logic of entity collapse comports more cleanly with the nature of the data inconsistencies they aim to resolve, lending credence to leveraging modern corporate structures for disambiguation of large corporate data.

A. Estimating Precision, Recall, and F1 Scores with Rare Events Regression Model Predictions

To estimate the value of these classifiers in collectively identifying accurate matches in the ground truth sample, we utilized the predicted probabilities of dyadic collapse from the model in a test-retest process. This involved first sampling the positions in the ground truth sample which had a collapse and those which did not, dividing each population on into 80% and 20% of their constituent adjacency matrix positions. We then used the 80% data to build a similar model to that above, and used its beta values to estimate predicted probabilities of dyad collapse in the remaining 20% of the data. We repeated this process 10 times to collect the average precision, recall, and F1 scores of the test model on the predicted collapses. The average and standard deviations of these iterations serve to indicate a true distribution for these parameters, presented in Table III.

Broadly, these results indicate significantly high efficacy of the classifiers through the rare events regression model predictions. The average precision is slightly lower than 0.98, indicating a high level of accuracy in the results presented by the classifiers. Furthermore, the predicted probabilities suggest over 10% of the ground truth dyad collapses. Importantly, these are the predictions of the classifiers when independently applied to the data without pre-processing of other classifiers. Cumulative applications may serve to collect an even higher proportion of true matches. Given the efficacy of these classifiers, it is important to test their effects on observed networks, in both topology and actor-level measure distributions.

TABLE III. REGRESSION-BASED PRECISION, RECALL, & F1 SCORES

N=10	***Precision***	***Recall***	***F1***
Average	0.9791	0.1160	0.2075
Standard Deviation	0.0015	0.0031	0.0049

VI. NETWORK EFFECTS OF ENTITY RESOLUTION

Finally, it is important to report on the observable implications of these classifiers for accurate feature extraction from the BoardEx employment network. Given that these classifiers impact entity count and frequency within the data so significantly, it is important to gauge their relative impact on features extracted from our observed network data. Using the likelihood of all pairwise matches given by the rare events model above, we iteratively reconstructed the network including these entity compressions. First, we generated an edgelist of entity merges with their predicted probabilities using the regression model. Then, we iteratively sampled the binomial distribution with these probabilities, and used the binary to dictate a collapse or non-collapse of the given entity dyad within the BoardEx employment network. Finally, we collected the relevant topology measures, and repeated the above process 100 times. These sampled topology scores allowed for an effective bootstrap of their average, to better gauge the distance of the network data without resolution from its resolved network. In the sections below we present the effects of the random regression-based simulations on two dimensions of network measures: topology and actor-level distributions. Regarding topology, we examine effects on density and degree centralization. We also examine the effect of these classifiers on actor degree distributions within the simulated and untreated networks.

A. Density

Density was the network feature most affected by these cleaning mechanisms. This is not an intuitive conclusion; it is quite possible that this figure could have diminished if a great number of ties existed only between artificially distinct nodes. However, as shown in Figure 2, the opposite is the case; the untreated network tends to underestimate density by about 2 standard deviations from the simulations mean. This follows from the large number of nodes reduced through these resolution techniques, and suggests that the true network is denser than the untreated data.

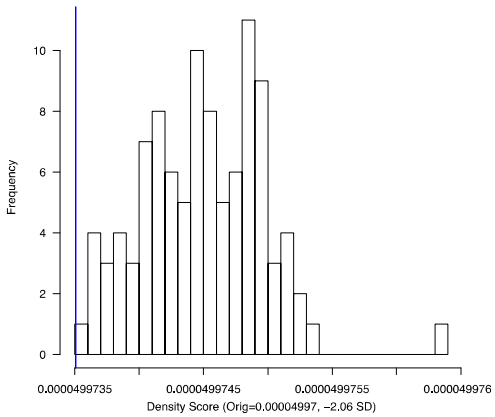


Figure 1. Histogram of simulated density scores, blue bar untreated.

This measurement error has social implications for researchers aiming to use BoardEx data. This error understates the true propensity of nodes to interact with one another as a function of their artificial duplicity. In measuring relevant employment phenomena, such as how often firms (of similar or dissimilar types) share employees, researchers may mistakenly understate the phenomenon and offer inaccurate results. Suggestions made regarding the causes or consequences of these phenomena will, by extension, be inaccurate and reflective of the reality suggested by the data, not the reality which the researcher strives to understand and describe.

B. *Degree Centralization*

While marginal, there was a slight tendency for the untreated employment network to underestimate network features of degree centralization among actors, as shown in Figure 3. Standing 0.37 standard deviations to the left of the simulation means, the untreated network reports lower than valid centralization among nodes as a function of their artificial duplicity. Once collapsed, central nodes become yet more central to a social network, and when artificially distinct their centrality is artificially deflated and dispersed across several actors.

There are several social implications of this feature extraction error. It stands to minimize researchers' ability to effectively identify hierarchy in networked systems. In substantive cases this can dampen research projects on corporate monopoly, tax practices, or any other number of corporate phenomena. Resolving artificial distinctions among nodes contributes to the reduction of this tendency to underestimate degree centralization in observed network data, thereby improving the accuracy of researchers relevant claims regarding the social world under study.

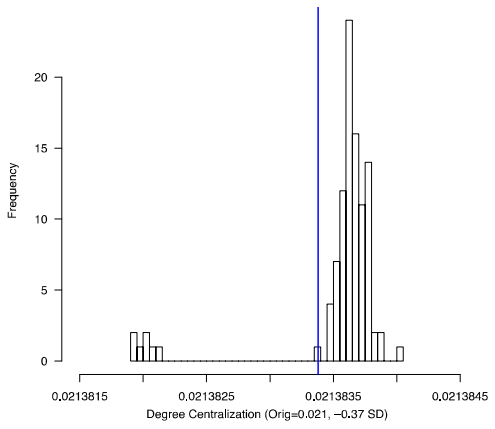


Figure 2. Histogram of simulated degree centralization scores, blue bar indicating untreated network degree centralization.

C. *Actor Degree Distribution*

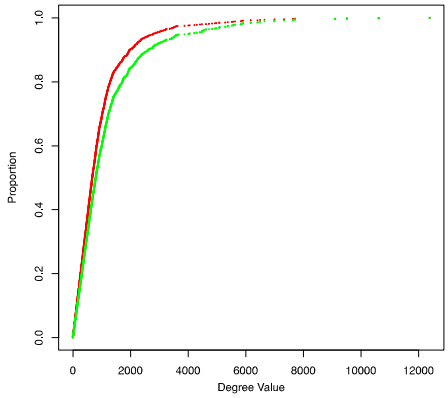


Figure 3. Cumulative degree distributions of untreated (red) and simulated (green) networks' actor degree centrality distributions.

Finally, the degree distribution of the untreated network exhibits a sharper curve than the treated networks. Especially between the degree values of 1,000 and 6,000, the untreated network over-reports the proportion of nodes with that many ties, an important feature of network data. As shown in Figure 4, this tendency is similar to that of degree centralization, namely its underestimation of nodes with very high degree scores. The social implications for this measurement error are also similar; researchers will be less likely to correctly identify overly central nodes in a corporate network when their entity names are artificially distinguished.

VII. CONCLUSION

The importance of accurate network data features for sound social science inference cannot be understated. Here we present a novel means of addressing issues in feature extraction resulting from data inconsistencies by leveraging the relational structure of a network dataset to collapse nodes which are truly indistinct. We integrate external relevant datasets (such as corporate hierarchies) and knowledge of syntactic modifiers which preclude exact computer matching (such as suffixes and corporate listing dates) to both clean and enrich the relational BoardEx data and improve the accuracy of extracted features. Importantly, we leverage the social structure of the modern corporation to effectively clean its multiple incarnations in big data.

We demonstrated that these methods each serve as successful classifiers, individually yielding relatively high precision and recall scores on a hand-coded ground truth sample of entities. We demonstrated with a rare events regression model the independent efficacy of each classifier. Finally, we demonstrate that there is an observable effect on extracted network features when individuals fail to account for these entity inconsistencies, yielding measurement errors ranging in standard deviations from the simulation means.

Broadly, it is clear that researchers paint only a partial picture of social phenomena if failing to account for these inconsistencies. This has implications both for existing findings and future work using large corporate data. First, while the error is small, it leans in a consistent manner in our test dataset. Failing to account for artificially distinct entities can underestimate network features of density and degree centralization, and overestimates cumulative proportions of nodes by their degree centrality scores. It is not unlikely that other such topological and distributional measures are impacted by these inconsistencies. Furthermore, entity duplicity can obfuscate significance in linear models used to evaluate all kinds of corporate phenomena. Second, future work should build on the relational approach offered here. This involves tailoring the relational logic to the given research question, and not all classifiers here may be necessary or appropriate. Further, these classifiers should be tested with other corporate datasets comparable to BoardEx in substantive or structural manner, in order to better estimate their relative efficacy more generally. Finally, researchers can experiment with the effect of various kinds and combinations of external data integration in testing classifiers to more clearly gauge their success as compared to the corporate reality they reflect.

ACKNOWLEDGMENT

We are grateful to Elisabeth Sullivan-Hasson, Matthew Winn and Piotr Wlodkowski for research assistance on this paper. James Heilman contributed significantly on elements throughout the project but not on this paper. We also gratefully acknowledge the financial support of the Russell Sage Foundation grant #83-15-13.

REFERENCES

[1] Benjelloun, Omar, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. "Swoosh: a generic approach to entity resolution." The VLDB Journal—The International Journal on Very Large Data Bases 18, no. 1 (2009): 255-276

[2] Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008) 'Fast Unfolding of Communities in Large Networks', *Journal of Statistical Mechanics: Theory and Experiment*, 10, doi:10.1088/1742-5468/2008/10/P10008

[3] Brizan, David Guy, and Abdullah Uz Tansel. "A. survey of entity resolution and record linkage methodologies." Communications of the IIMA 6, no. 3 (2006): 5.

[4] Bryan, Dick, Rafferty, Michael and Wigan, Duncan. 2017. "Capital Unchained: Finance, Intangible Assets and the Double Life of Capital in the Offshore World", *Review of International Political Economy* 24(1): 56-86.

[5] Choirat, Christine, D'Orazio, Vito, Honaker, James, Idris, Muhammed and McGrath, Jennifer. 2016. "Interpreting Zelig: Everyone's Statistical Software", mimeo, 9 August.

[6] Corporate Affiliations. [Electronic Resource]. New Providence, N.J. : LexisNexis Group., n.d. EBSCOhost

[7] Cuxac, Pascal, Jean-Charles Lamirel, and Valérie Bonvallot. "Efficient supervised and semi-supervised approaches for affiliations disambiguation." Scientometrics 97, no. 1 (2013): 47-58.

[8] Dey, Lipika, Ishan Verma, Arpit Khurdiya, and Sameera Bharadwaja. "A framework to integrate unstructured and structured data for enterprise analytics." In Information Fusion (FUSION), 2013 16th International Conference on, pp. 1988-1995. IEEE, 2013

[9] Dubrawski, Artur, Kyle Miller, Matthew Barnes, Benedikt Boecking, and Emily Kennedy. "Leveraging publicly available data to discern patterns of human-trafficking activity." Journal of Human Trafficking 1, no. 1 (2015): 65-85.

[10] Elmagarmid, Ahmed K., Panagiotis G. Ipeirotis, and Vassilios S. Verykios. "Duplicate record detection: A survey." IEEE Transactions on knowledge and data engineering 19, no. 1 (2007): 1-16.

[11] Fortunato, Santo. "Community detection in graphs." Physics reports 486, no. 3 (2010): 75-174.

[12] García-Moya, Lisette, Shahad Kudama, María José Aramburu, and Rafael Berlanga. "Storing and analysing voice of the market data in the corporate data warehouse." Information Systems Frontiers 15, no. 3 (2013): 331-349.

[13] Gonzáles-Bailon, Sandra, Jennings, Will and Lodge, Martin. 2013. "Politics in the Boardroom: Corporate Pay, Networks and Recruitment of Former Parliamentarians, Ministers and Civil Servants in Britain", *Political Studies,* 61(4): 850-873.

[14] Gurney, Thomas, Edwin Horlings, and Peter Van Den Besselaar. "Author disambiguation using multi-aspect similarity indicators." Scientometrics 91, no. 2 (2012): 435-449.

[15] Heemskerk, Eelke, Young, Kevin, Takes, Frank, Cronin, Bruce, Garcia-Bernardo, Javier, Popov, Vladimir Henriksen, Lasse Folke and Winecoff, W. Kindred. 2017. "The Promises and Perils of Using Big Data in the Study of Corporate Networks", *Global Networks*, 18(1), in press

[16] Köpcke, Hanna, Andreas Thor, and Erhard Rahm. "Evaluation of entity resolution approaches on real-world match problems." Proceedings of the VLDB Endowment 3, no. 1-2 (2010): 484-493.

[17] Lalanne, Marie, and Paul Seabright. "The old boy network: Gender differences in the impact of social networks on remuneration in top executive jobs." (2011).

[18] Liu, Wanli, Rezarta Islamaj Doğan, Sun Kim, Donald C. Comeau, Won Kim, Lana Yeganova, Zhiyong Lu, and W. John Wilbur. "Author name disambiguation for PubMed." Journal of the Association for Information Science and Technology 65, no. 4 (2014): 765-781

[19] Londhe, Nikhil, Vishrawas Gopalakrishnan, Aidong Zhang, Hung Q. Ngo, and Rohini Srihari. "Matching titles with cross title web-search enrichment and community detection." Proceedings of the VLDB Endowment 7, no. 12 (2014): 1167-1178.

[20] Robinson, David. "The Use of Reference Graphs in the Entity Resolution of Criminal Networks." In Pacific-Asia Workshop on Intelligence and Security Informatics, pp. 3-18. Springer,, 2016.

[21] Shahri, Hamid Haidarian, Galileo Namata, Saket Navlakha, Amol Deshpande, and Nick Roussopoulos. "A graph-based approach to vehicle tracking in traffic camera video streams." In Proceedings of the 4th workshop on Data management for sensor networks: in conjunction with 33rd International Conference on Very Large Data Bases, pp. 19-24. ACM, 2007.

[22] Smalheiser, Neil R., and Vetle I. Torvik. "Author name disambiguation." Annual review of information science and technology 43, no. 1 (2009): 1-43.

[23] Szekely, Pedro, Craig A. Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor et al. "Building and using a knowledge graph to combat human trafficking." In International Semantic Web Conference, pp. 205-221. Springer, Cham, 2015.

[24] Yujian, Li, and Liu Bo. "A normalized Levenshtein distance metric." IEEE transactions on pattern analysis and machine intelligence29, no. 6 (2007): 1091-1095.

[25] Zhao, Jianyu, Peng Wang, and Kai Huang. "A semi-supervised approach for author disambiguation in KDD CUP 2013." Proceedings of the 2013 KDD Cup 2013 Workshop, p. 10. ACM, 2013