Leveraging Web Data to Monitor Changes in Corporate-Government Interlocks in India

Anirban Sen, A. Agarwal, Aditya Guru, A. Choudhuri, G. Singh, Imran Mohammed, J. Goyal, K. Mittal, Manpreet Singh, Mridul Goel, S. Gupta, S. Pathak, Varuni Madapur, Aaditeshwar Seth
IIT Delhi
{anirban,aseth}@cse.iitd.ac.in

ABSTRACT

Corporate executives who are linked to politicians or administrative officials, or family members of public officials with links to corporate organizations, are known to build an interlocking social network that becomes a power structure of highly influential entities. Such power structures often lead to an inequitable distribution of resources and manipulation of policies. A deeper look at this power structure and the constituent interlocks can provide users with valuable insights on these influential connections, and eventually, on the shaping of socio-economic outcomes by the interlocked political economy. In this paper, we describe the design of a platform to empirically monitor the degree of corporate-government interlock in India over time, by making use of publicly available data on the web. We find that the interlock has strengthened over the last decade, and we report the kind of interconnections and structural changes that have happened during this time. We also describe the design of an application to present news articles about an event or topic alongside the interconnection network of entities referred to in the news articles, to help users get a quick view of the main actors involved in the event. We find that this news search application is able to highlight several interconnections between prominent entities in an event, which had not been reported in the media. Overall, we find it relevant to build a technology platform which can help researchers and journalists to monitor the extent of interlocks between powerful stakeholders in the corporate and government spheres.

CCS CONCEPTS

• **Information systems** → **Information systems applications**; • **Social and professional topics** → *Computing / technology policy*; • **Applied computing** → Computers in other domains;

KEYWORDS

Media analysis, power elite, topic modeling, sentiment analysis

1 INTRODUCTION

Corporate executives who enter politics or government administration, or family members of public officials with links to corporate organizations, are known to build an interlocking social network that becomes a power structure of highly influential entities [19]. This power structure often leads to a bidirectional flow of favors between the corporations and political entities [9], and it can influence policy formulation or manifest itself in cronyism [2, 20]. Cronyism, as defined by Begley et al. [3] is 'preferential treatment shown to old friends and associates without regard to their qualifications'. With diminished competition and increased corruption, cronyism also impacts how well the state is able to redistribute wealth and create policies for social welfare [8]. Such negative manifestations of corporate-government interlocks have been evident in India [12, 31]. Given the growing wealth inequalities and fears of elite capture of public institutions to influence policy [25], it is therefore important to develop indicators and tools to monitor corporate-government interlocks. Having information about changes in the interlock structure can give researchers and journalists valuable context to interpret economic and policy changes.

Piece-meal analysis of interlocks is done by journalists in an investigative manner, but no comprehensive automated mechanisms exist currently. In this direction, we have built a technology platform which gathers data from the web to be able to monitor changes in corporate-government interlocks over time. One of our key contributions has been to use this platform in India to gather and curate data about 19295 central and state level politicians, 11531 bureaucrats, 64155 companies, and 111105 managers, over the period of 2004 to present, by crawling numerous structured and unstructured web data sources. Doing this required several technical innovations, including a system to keep track of multiple versions of the social network graph as it is augmented with new data, entity resolution algorithms that can improve their resolution accuracy over time by gaining more and more context information about entities, and graph analysis methods to mine the data. We have codified four kinds of interlock *patterns* on this data, to study the evolution of networks of *corporate connected politicians, corporate connected bureaucrats, politically connected firms, and politically connected managers*. We have developed and validated methods to rank entities according to each of these four patterns, and used these rank scores to build an indicator of corporate-government interlock as an empirical measure of the extent of this interlock in the country. Finally, we have built a novel news browsing interface which operates on a corpus of more than 4 million news articles gathered from seven leading news dailies in India from 2011 to present, and given a query topic or event, it extracts a concise subgraph of important entities involved in the event and ranks news articles based on their coverage of these entities – this provides a tight *nugget* of information to help users get a quick view of the main actors in the event and read about their role in the event. This too is a novel technical contribution in

COMPASS '18, June 20–22, 2018, Menlo Park and San Jose, CA, USA

ACM ISBN 978-1-4503-5816-3/18/06…$15.00
https://doi.org/10.1145/3209811.3209822

news presentation, by augmenting news articles with information about the interconnection network of entities referred to in the news articles, to help users interpret the news more easily.

Our analysis indicates that the corporate-government interlock in India has increased over time. We find that the number of interlocking corporate and government actors has increased, especially due to bureaucrats taking up directorships and high ranking positions in companies after their retirement. Further, the corporate network has become more concentrated due to new ownership links and shared directors between companies. We use this analysis and demonstration of our news browsing interface to invite researchers and journalists to use the platform and make the insights more accessible to people. Corporate-government interlocks are an important filter to interpret the political economy of policy making and cronyism [11], and is increasingly important as growing wealth inequality makes it easier for powerful actors to leverage these relationships to their advantage [25]. Our platform can be easily adapted for use in other countries, or even trans-nationally, and enable people to monitor the players controlling the path of their economy towards more equitable development. In this paper, we mostly describe the design of the platform and present an initial analysis and case-study of a corruption scam in the telecom sector; we are now working on systematically analyzing multiple events using the platform to be able to find out more recurring patterns and strengthen our analysis.

2 RELATED WORK

We describe related work along four streams we are straddling, namely the reasons to study corporate-government interlocks, other open data initiatives in this domain, methods for entity resolution in structured and unstructured data, and news browsing interfaces to use social network data for new ways of presenting news.

2.1 Corporate-government interlocks

Corporate-government interlocks are an indicator of strong collaboration between the state and corporate sectors, which can lead to the formation of influential power structures. In the book *The Power Elite* [19], the author critiques the network of power in the United States which has significantly shaped the economy and government. On similar lines, in *The Price of Inequality* [25], the author talks about income inequality that results from these networks through rent-seeking and bidirectional flows of benefits between the corporate and political domains, and a positive feedback loop which sets in because increased inequality makes it easier for influential people to leverage their networks for personal gain. There have been ample studies in this direction, we list here only some of them.

Flow of favor from the political to the corporate domain: Mian et al. [18] show in their work from Pakistan that bigger, politically connected firms have access to loans of much higher amounts from government banks as compared to unconnected firms. Another study by Mara Facio [9] carries out its analysis across several countries and observes sharp increases in stock prices of companies whenever they form political connections. Some studies like [10, 15] from Indonesia and Malaysia, similarly show that the fortune of politically connected firms is highly dependent on the fortunes of the politicians they are connected to.

Flow of favor from the corporate to the political domain: Bertrand et al. [4] find that in France, in the wake of municipal elections, firms with politically connected CEOs see a sudden rise in employment rates (intended to glorify the concerned politician's image), and a drop in job destruction (firing). In the Indian context, Sandip Sukhtankar [26] provides evidence that during election years, politically connected sugar mills pay lower prices of sugarcane to farmers – this saving by the sugar mills is used to fund election campaigning, and is later passed on to farmers through waivers and other public policies if the politician wins.

While the works discussed above provide evidence of specific cases where corporate-government interlocks manifest themselves, there is no easy way or a service where researchers and journalists can analyse interlocks, identify curious patterns, and investigate these red-flags in more detail. Our platform is an effort in this direction, to make it easier to identify undesirable outcomes from corporate-government interlocks. To the best of our knowledge, ours is a novel contribution in this area, and also in several technical innovations we discuss later in the paper.

2.2 Other platforms providing corporate and political data

There are a few open data initiatives that provide data on corporate and political entities at the global level. OpenCorporates [27] shares data on corporate entities from many countries, and has been used to uncover several trans-national corporate ownership networks. Littlesis [13] shows connections between powerful people and organizations by tracking the key relationships of politicians, business-people, financiers, and their affiliated institutions. Unlike these platforms however, our goal is not just to collect and host such data but to build specific applications and provide inbuilt analysis tools which allow easy exploration of the data to obtain insights. Further, data about India on these platforms is not as rich as the dataset we have put together by integrating multiple data sources together, and the data cleaning and curation tools we have developed can be used to enrich the data on these platforms as well.

2.3 News presentation

Enriching the context of news media is an important research area, to provide useful background knowledge to people to help them interpret news events. Shahaf et al. in their paper [22] propose a technique of connecting entities occurring in the news articles to help users navigate the news topic. Similar work by Lu et al. [17] proposes a method of video summarization by selecting a chain of video subshots showing essential events from a long input video. The GDELT Project [16] monitors the broadcast, print, and web news globally in different languages and performs an analysis to identify the people, locations, organizations, themes, sources, emotions, quotes, and events, covered in the news. Our work with the design of news browsing interfaces is on similar lines to make it easier for users to understand about the main entities involved in the news, but differs in its approach of context enrichment by referring to an external knowledge base of a named entity social network of prominent corporate and government actors, to help people interpret the news articles more easily. Stasko et al. developed a system named *Jigsaw* [24], which is similar to our system in terms of its goal to capture the representation of entities and their interconnections between them. However, Jigsaw only uses text documents for discovery of

entities and co-occurrence based interconnections, while our system uses both textual news articles and a social network formed from structured data sources. Moreover, our system is specifically focused on exploring corporate-government interlocks unlike Jigsaw, which works in a general domain.

2.4 Entity resolution

Entity resolution (ER) is an important component in our system as it is needed to construct the social network of various entities by combining data from multiple structured and unstructured web-based data sources. While initial studies in ER [6, 7] only used partial string matches and distance measures to resolve named entities, research later moved on to context based or secondary-source based resolution where additional co-occurring information was used for ER [5, 23]. While our work does not make use of the aforementioned techniques directly, we borrow ideas to use context information such as co-occurring entities in the social network, properties of entities, or co-occurring words in news articles, and use that to incrementally learn new context information about the entities. Our ER approach achieves a satisfactory performance using such contextual resolution, and improves with time as more and more context information is gathered about the entities.

3 DATA

As explained in the previous section, the two infrastructure components of the graph store and the media database, obtain their data from publicly available online sources. In the following sections, we provide the details of this data.

3.1 Social network data

We have painstakingly put together the social network data from multiple structured and unstructured online sources over a period of time. This data now includes information about politicians, bureaucrats (officers of the Indian Administrative Services), firms and their subsidiaries, business-persons (board of directors (BoDs) and executive management of the firms), and family members of the business persons and politicians. Table 1 shows the details. We are able to obtain data of politicians who contested the national elections from 2004, all chief ministers of different states over time, retired and current bureaucrats in government service, all government departments and ministries at the center and state level, and a snow-balled network of companies and their board of directors starting with the approximate 5000 public listed companies in India. The snow-balling of the company network was done up to a depth of 3 to include subsidiaries of companies and companies connected through co-occurring members in their board of directors. This network therefore contains the largest companies and most important politicians and business-persons, in line with our goal to be able to examine corporate-government interlocks among the big corporate houses and prominent politicians. The data collection was a tedious task. Accessibility was a key challenge since a lot of relevant data was available in PDF documents in formats that were different each year, and our parsers had to be modified to extract the data[1]. Non-uniformity of web interfaces was another hurdle; some websites provided easily accessible dumps while others were interactive and required automated query generation to obtain data. Further, data had to be clubbed together from multiple sources. For example, data about the subsidiaries of public companies was available on the stockmarket websites, but data about the board of directors of these companies was available on a different website, and data about executive managers on yet another website. All this led to a significant challenge in entity resolution (ER) especially when unique IDs were not given, and names of people or companies were spelt differently in different places. We briefly describe this problem of ER next.

Entity resolution in the social network: We have several types of entities in the network: People (including politicians, bureaucrats, business-persons, and their relatives), companies, constituencies, states, ministries, departments, and political parties. The ER process for these entities works in two steps: (i) matching using the context data for entities, and (ii) further filtering using string and phonetics based similarity measures on the entity names.

Different pieces of context information are used in different cases based on the available data. To resolve newly crawled *politicians* with existing politicians, we use their political parties and political titles (like prime minister, chief minister, etc.) as the context information. To resolve them with business-persons and bureaucrats, we use the context information available in Wikipedia's *Category* section of the entities' pages[2]. *Bureaucrats* are resolved against business-persons using their date of birth and names. Finally, *Business-persons* are resolved among themselves using the *company identification numbers* (CINs) of the companies with which they are associated. To resolve *companies* with each other, data from some sources was straightforward to match based on the CINs, and in other cases context information was used such as the registered location of the company.

After context based matching, sometimes in cases where a precise identification could not be done, the candidate list of matching entities were resolved using string+phonetics based approaches on an experimentally decided threshold of similarity. We measured the performance of our ER approach based on randomly selected 2000 person entities, with equal number of positive and negative examples[3] annotated by two of the co-authors of this paper. We obtained precision and recall[4] values of 95% and 97.4% respectively (apart from companies (which could be mostly resolved by CINs), there are very few non-person entities. Hence, the ER performance for them were nearly 100%.), and managed to build a social network graph with relations of the types: *politician—belongsto—political party (20951), subsidiary—belongsto—company (453)*, etc. Figure 1 provides a high level view of the social network graph data where these relationships are shown. Although our currently obtained relationships are all extracted from the structured social network data, we are currently augmenting this with relationships extracted from news articles using NLP and *association rule mining based* techniques [1, 35]. However, we observe that most relationships obtained

[1] We could not use procurement documents for data collection as most of them exist in the form of scanned, handwritten digital copies, which are difficult to parse.

[2] For example, if a politician has also been a bureaucrat at some point before her political career, usually *Indian Administrative Service* is mentioned in the category section of her Wikipedia page.

[3] Positive: new entities that can be merged with existing entities, Negative: entities that do not match and hence cannot be merged with existing entities.

[4] Throughout the paper, we consider correctly resolved entities as TP, correctly unresolved entities as TN, incorrectly resolved entities as FP, and incorrectly unresolved entities as FN.

Type of Entity	Count	Dataset	Attributes	Timestamps	Sources
Politicians	19295	Lok Sabha	Name, constituencies, political party	2004-09, 2009-14, 2014-till date	www.indiavotes.com
		Rajya Sabha	Name, constituencies, political party	2014-till date	www.wikipedia.org
		Prime/chief ministers	Name, constituencies, political party	All till date	www.wikipedia.org
Bureaucrats	11531	IAS Officers	DoB, department, districts or locations served, educational qual., training details	1961-till date	www.persmin.gov.in
Business-persons	111105	Directors (managers)	Director ID (DIN), name	1961-till date	www.mca.gov.in
Firms	64155	Listed firms	Income, subsidiaries	1980-till date	www.bseindia.com
		Listed/non-listed firms	Company ID (CIN), name, headquarters, authorization capital, date of incorporation	1980-till date	www.mca.gov.in
Departments	1565		Name, ID	All till date	www.wikipedia.org
States	36	Name	Current		www.wikipedia.org
Family information	71	For politician	Name	All-till date	www.wikipedia.org
	222	For managers	Name		www.wikipedia.org

Table 1: Overview of web data collected for the social network

from mass media data are corresponding to important politicians and there is sparsity of information on other entities like bureaucrats and companies. This is the primary reason that we use our tediously collected structured data to extract relationships of varied types, which is one of our key contributions.

3.2 Media data

Media data is crawled on a daily basis from some of the most popular Indian news sources, *The Hindu, The Times of India, Indian Express, The New Indian Express, Telegraph, Deccan Herald and Hindustan Times*, and archives were used to build a corpus of news articles since 2011. The data is stored at three levels: article URLs and their meta-data, article text, and entities extracted from the article text.

Entities are extracted from the article text using the *OpenCalais* service [21], which provides entities of type Person, Company, Organization, City, Province, and Country from an article. Along with the entities, OpenCalais also provides additional context attributes like the type of entity, its standard name, and some other context information (in case of non-person entities)[5]. We also maintain a set of *aliases* for each resolved entity, which keeps getting enriched with standard names of the newer entities that are resolved with it.

Similar to the social network, an entity might occur in different forms in the media data too, leading to the need for ER. Since news articles are crawled continuously, we maintain a set of *resolved entities* which have been successfully resolved so far, and keep augmenting it as more news articles are crawled and throw up additional entities to be resolved. On encountering a new unresolved entity, ER within media data follows two steps: (a) find the top ten candidate

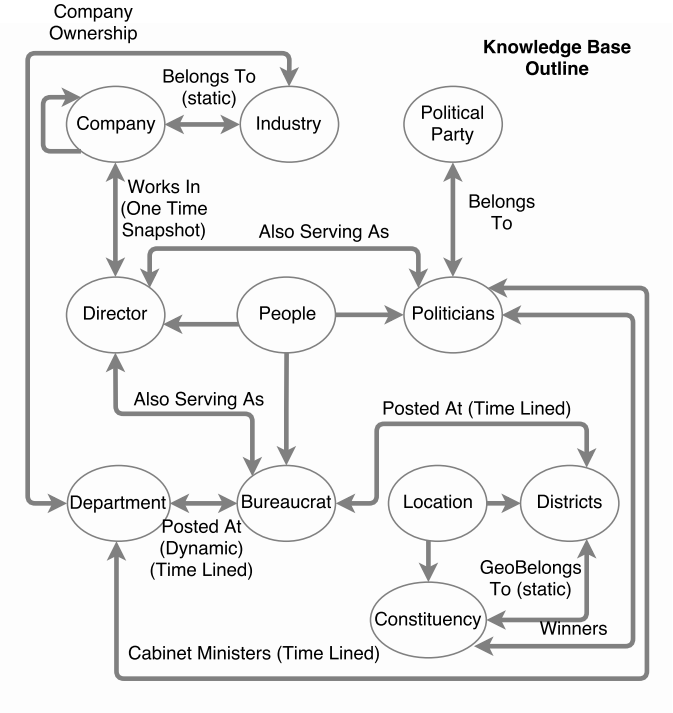


Figure 1: High level overview of the data in the graph social network graph database: the timed and untimed (static) relations are shown in edge labels.

[5]For example, latitude, longitude, state, and country information are returned in case of cities.

Interval	Person		Non-person	
	Precision%	Recall%	Precision%	Recall%
2011-12	87.5	93.9	88.6	94.59
2012-13	83.52	95.94	81.17	97.18
2013-14	89.41	98.7	91.02	98.61
2014-15	91.95	95.23	86.58	100
2014-16	95.18	92.94	89.02	97.33
2016-17	97.61	96.47	93.82	96.2

Table 2: Precision and recall percentages of ER within the media database for the person and non-person (Object) entities: there is an overall improvement due to context enhancement over time.

entities from the resolved set of entities based on partial matching (match of at least one word) of their standard names, aliases, and context; and then (b) further filter these top ten entities to obtain a set of best matches, using string+phonetics based distance measures applied on standard names and context of entities, based on experimentally set similarity thresholds (a combination of Jaro-Winkler similarity and Levenstein distance, along with substring and abbreviation matching). The ER process uses contextual attributes returned by OpenCalais, and we follow a strategy of merging this context information together for entities which are successfully matched with each other. This improves the accuracy of the resolver over time as it gains more and more context information for each resolved entity. If any of these steps fail, we make a new entry in the resolved entity set. To evaluate the performance of ER within the media data, we collected 100 random entities from the last 50 news articles for every year from 2011 to 2017. For each time period, the 100 unresolved entities from these 50 articles are resolved against all of the remaining resolved entities whose context information has been enriched with data from the previous years combined. For example, entities from the last 50 articles within the 2012-13 period are resolved with the remaining resolved entities during 2011-12 and 2012-13. Table 2 shows the precision and recall values for each of these sets of last 100 entities. Although the values do not exhibit a strict monotonic improvement with increasing context, we can see that there is a significant overall improvement from the first time period (2011-12) to the last (2016-17) in both of the types of entities (Person and Non-person).

The media entities thus resolved, are next also resolved against the entities in the social network (Neo4j graph store). This process of ER between the media and social network entities starts with alias matching, and further filters the result set through context matching (an approach similar to the ER approach for social network). We also use *associated entities* for context matching, which are entities co-occurring in any article in the media data, or neighbouring entities in the social network data, and it is able to improve the ER performance significantly. The performance of ER between the social network and media entities was also satisfactory. We randomly selected 50 person and 50 non-person entities and observed the precision and recall values as 93.18% and 95.35% respectively for persons, and 87.5% and 89.74% respectively for non-persons.

4 SYSTEM OVERVIEW

Our system consists of two infrastructure components: the social network graph store, and the news media crawler and article store. On top of these infrastructure components, we build multiple applications. Here we discuss the *interlock monitoring application*, which tracks changes in the corporate-government interlock patterns to find instances noteworthy of further investigation, and the *news nugget application*, which given a topic or event outputs the interconnection network of entities relevant to the topic along with important news articles about the topic. The block diagram in figure 2 provides a brief overview of our system architecture, which we now describe below.

The **social network graph** is formed out of data obtained from different web sources, and is stored in a Neo4j [29] graph database. We use Neo4j, since graph databases to store social network data are known to perform better than general relational databases for path extraction queries [30]. In Neo4j, entities are represented as nodes, the relationships between them are represented as edges, and both nodes and edges have multiple properties (attributes) attached to them. A set of **crawlers** periodically visit a series of pre-identified web sources to obtain data on politicians, directors, bureaucrats, firms, political constituencies, locations of the firms' registered offices, relatives of the politicians, and other relevant information from publicly available websites.

A **resolver** module extracts and resolves entities in the crawled data. Ambiguous resolutions (that the resolver was unable to resolve automatically) are flagged for manual moderation by humans. A web-based user interface is provided where moderators can look at candidate matches and select the correct matches. We are extending this interface to even build a crowd-sourcing platform where people can add relationships along with evidence of the relationships, and post it for verification by the moderators.

Users, especially researchers and journalists, may also want to trace the source of the information stored in the social network, and will need knowledge of where and when the underlying data was crawled. We have therefore built an infrastructure that uses the Neo4j graph database in conjunction with Mysql, with Mysql used to store provenance information about all updates made on the social network graph. This provenance meta-data logs all updates to properties of existing entities and relationships (or insertion of new entities, relationships, or properties), the time of update, the source, old value, and the new value. Storing this meta information in a separate database does not impose any overhead on Neo4j for graph operations or queries. This also enables a rudimentary versioning system in case users are interested to run an analysis by including only a limited set of sources; the meta information can be replayed to create new instances of a Neo4j graph store by including or excluding the desired web sources.

The **media module** continuously crawls all articles on a daily basis from seven popular newspapers in India. The article text is stored in a NOSQL document database, and the URLs along with their meta-data and identified named entities are stored in a relational database. Similar to the social network module, it contains a **crawler** and a **resolver**. The resolver resolves entities extracted from freshly crawled articles against the resolved entities identified in older articles. It also matches the resolved media entities with

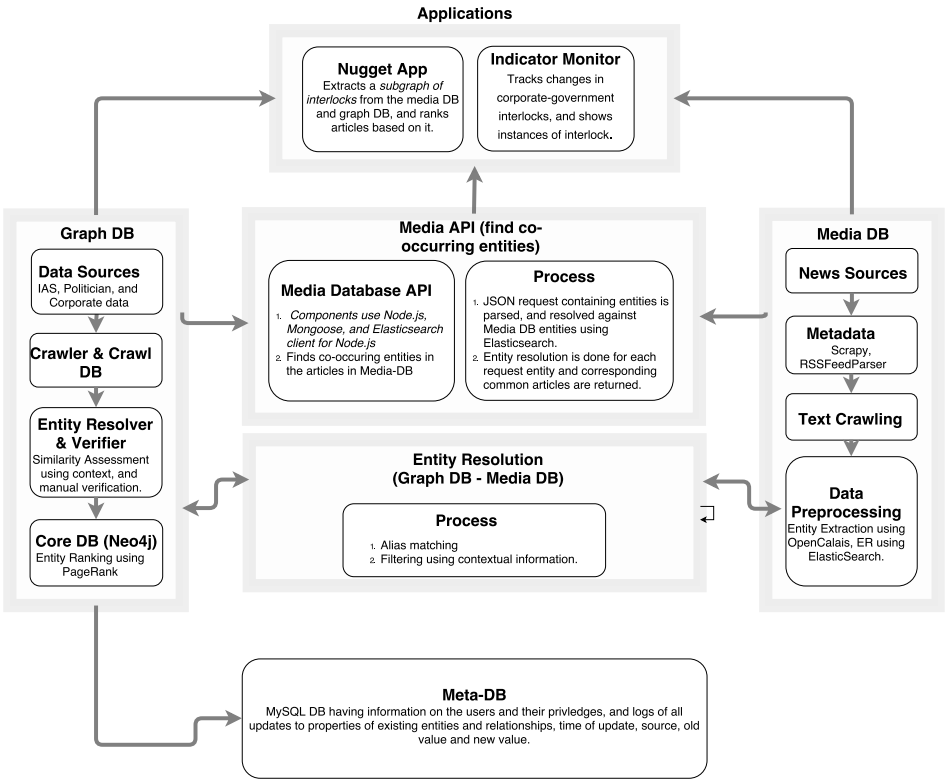


Figure 2: Overall architecture of the system

entities in the social network graph store. The number of entities present in the media data is of the order of millions, which makes ER within the media data a challenging task as each unresolved entity has to be compared pairwise with every resolved entity. We use an indexing solution called Elasticsearch [28] to make this efficient.

The **interlock monitoring** application then presents data about changes in corporate-government interlocks using an indicator we develop in this paper.

The **news nugget application** takes as input a search query for a topic or event, and extracts a 'nugget' of information drawn from the media database and graph database (social network) – it identifies the most mentioned and top-ranked entities mentioned in the event, extracts the interconnection paths between these entities to obtain a tight subgraph, and identifies the most explanatory news articles which maximally cover the subgraph.

5 COMPUTATION OF ENTITY SCORES

To distinguish important interlocks from not-so-important interlocks, we wanted to rank the entities in the social network graph on various criteria. In this section, we describe our method for this entity ranking. The computations are performed in an offline manner and can be triggered manually upon any significant data update in the social network graph.

We rank the entities based on their involvement in different types of patterns: *corporate connected politicians*, *corporate connected bureaucrats*, *politically connected firms*, and *politically connected managers*. The rankings are done through a series of PageRank computations for each pattern to obtain rank scores as a measure of an entity's involvement in instances of a pattern. We start with a global adjacency matrix (R), extracted from the social network graph described in the System Overview section, which contains all the entities and their relationships, we then compute R^4 to obtain a matrix which incorporates interconnection information over 4 hops, and then obtain sub-matrices of R^4 such as a company-to-company matrix, or a politician-to-company matrix, to obtain mutual associations between specific types of entities only. These association matrices are used for the PageRank computation described below, along with also computing a set of bias vectors which capture entity specific weights independent of the pattern being considered[6]. The

[6]As an example, the set of properties used to build the bias vector for bureaucrats are selected by using OLS regression analysis based on a hypothesis that considers their membership in the board of a company as a dependent variable, and includes independent variables such as the number of weeks spent in foreign training, their educational qualification, number of weeks spent in important departments, their designation, appointment in central ministries, and total tenure till date. We selected only those independent variables that had a p-value less than 0.001 as the final features to calculate the bias vectors.

Across different patterns, the following attributes are linearly combined to calculate the bias vectors for the entities:

(1) For companies, we use their authorization capital to calculate the bias vector E_c.
(2) For politicians, we use the number of times they were head ministers, cabinet ministers, or members of the parliament to calculate the bias vector E_p.
(3) For bureaucrats, a combination of time spent in important designations and departments (we consider pay-grade above a certain threshold as an important designation, and identify important departments as those classified according to [32] and [14].) is used to calculate the bias vector E_b.

steps described below are for the pattern of *corporate connected bureaucrats*. Entity scores for other patterns are obtained similarly:

(1) Apply PageRank on the company-to-company adjacency matrix (M_{cc}) obtained as a sub-matrix from R[4], with normalized authorization capital of the companies as the bias vector (E_c), to obtain a scoring of companies based on their corporate connectedness with other companies (C_n).

$$C_n = \lambda * M_{cc} * C_n + (1-\lambda) * E_c$$

(2) Multiply the bureaucrat-to-company adjacency matrix (M_{bc}) with C_n to get a scoring of bureaucrats based on their connectedness with companies (B_c).

$$B_c = M_{bc} * C_n$$

(3) Obtain the hybrid scoring of bureaucrats (E_h) by linearly combining their corporate based score vectors (B_c) with their bias vectors (E_b) containing information about their bureaucratic strength.

$$E_h = \beta * B_c + (1-\beta) * E_b$$

(4) Apply PageRank on the bureaucrat-to-bureaucrat adjacency matrix (M_{bb}), with the hybrid vector obtained in step 3 as a bias vector.

$$B_n = \lambda_1 * M_{bb} * B_n + (1-\lambda_1) * E_h$$

Thus, the final scoring of bureaucrats (B_n) is made to depend on their connections to companies, their bureaucratic strength, the strength of companies to which they are connected, and their connection to other bureaucrats. We call this rank score of an entity w.r.t. a pattern as the **entity-score** of the entity. The basic structure of our set of PageRank equations are the same for other patterns as well.

Validation of network computation: We next validate our method to check if it is able to appropriately rank entities heavily involved in the pattern considered. Using the entity scores for the *corporate connected bureaucrats* pattern explained above, we do this by examining how many of the top ranked bureaucrats have corporate connections within 4 hops in their neighborhood (obtained from the R matrix).We identified the top 114 bureaucrats out of 11531 bureaucrats in our dataset, by picking a threshold as the knee point in the entity score distribution. Out of these top bureaucrats, we found that 108 of them had corporate connections within 4 hops, giving us a precision[7] of 94.74%. The high value of precision indicates that our heuristic is indeed able to capture entities highly involved in the pattern, i.e., it is assigning high ranks to bureaucrats who are actually corporate connected in their neighborhood[8].We were not able to validate in the same way as other patterns for linkages of politicians due to less data on interlocks of politicians currently, and are looking at mechanisms to augment this through crowd-sourcing and media data. We find that among the most talked about entities in mass media in the context of *Demonetization* [33] (a recent policy event to crack-down upon black money and corruption), more than 90% of the entities are included in our dataset. Those not included are advisors or members of special committees, and we are planning to augment our dataset with these entities in the future.

[7]True positive: number of top ranked bureaucrats with corporate connections within 4 hops, false positive: number of top ranked bureaucrats with no corporate connections within 4 hops.

[8]The reason we could not report the recall here is that Neo4j takes a very long time to return results of multiple hop queries for even a small number of entities. It was therefore infeasible to run the query on our set of around 12K bureaucrats.

6 THE INTERLOCK MONITORING APPLICATION

The goal of the *interlock monitoring application* is to make it easier for users to observe changes in the corporate-government interlock over time, and discover curious patterns worthy of deeper investigation. To do this, we develop an indicator to quantify the strength of the corporate-government interlock, and enable closer investigation of the reasons behind changes in the indicator values over time. In the rest of the section, we describe our analysis of these changes to motivate the relevance of our work. Our main finding is that the interlocks have strengthened over the years, most prominently through appointments of bureaucrats in the corporate sector after their retirement, and an increasing concentration in the corporate sector through denser company ownership networks and interlocks of shared directors especially across large companies.

We define the *interlocking network* as the bridge edges (extracted from the R matrix) that connect any entity on the government side (politicians or bureaucrats) with any entity on the corporate side (companies, or managers, or members of the BoD). This includes edges like *politician—worksin—company* and *politician—relatedto—manager*. The indicator is computed as the sum of the scores of the bridge edges in the interlocking network:

$$I_{cp} = \sum_{i=1}^{|E|} edgescore(e_i)$$

where *edgescore* is defined as the product of the *entityscores* of the end-points of the bridge edges (e_i), and $|E|$ is the number of bridge edges. The *edgescore* of a bridge edge will thus have a high value only if both of its endpoints have high value of *entityscores*. This essentially models the probability of the interlocking network to exercise its influence.

Change in interlocks over time: Since we are aware of the timestamps when most edges formed, we are able to show in Figure 3 the change in index value over time. We consider time periods between the general election years of India (2004, 2009, and 2014) as entry of new legislators during these election years will lead to changes in the interlocking scenario. As we can see, the index of interlock shows a monotonic increase with time, indicative of an increase in corporate-government interlocks.

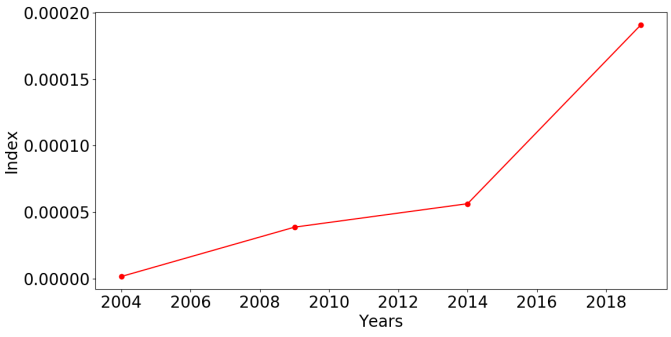


Figure 3: Change in index of corporate-government overlap over time (2004-till date)

Bridges	Untimed	Before 2004	2004-2008	2009-2013	2014-2018
POL-COM	3	12	10	8	8
POL-BoD	8	6	6	1	1
IAS-COM	258	22	109	249	353
IAS-COM (Govt./Public)	701	4	79	419	692

Table 3: Count of bridge edges over time (untimed edges were considered for calculations across all time periods). POL, COM, BoD, IAS stand for politicians, companies, directors, and bureaucrats respectively. The Govt/Public links are for appointments of bureaucrats in state owned companies, and are not considered in the calculations.

Causes behind increase in the interlocks: There can be two structural reasons behind the increase observed in the indicator values: (a) the formation of new bridges over time, and (b) increase in scores of the existing bridges due to either an increase in concentration in the government network or in the corporate network. To investigate this further, in table 3, we show how the number of bridge edges changes over time.

We can see that the number of bridge edges corresponding to especially the *corporate connected bureaucrats* pattern significantly increases with time. This observation is validated by sources [34], which mention how bureaucrats are increasingly taking up corporate positions after their retirement. One limitation that occurs due to the lack of data in our work is that due to missing timestamp data on some edges, we are not able to identify the direction of interlock formation in all cases. For example, whether a politician joined a company as a BoD (director), or whether a BoD entered politics. We plan to address these limitations by crowd-sourcing more data.

To study the second reason of an increase in concentration of the government and corporate networks, figure 4 shows the CDF of the degrees of the interlocking bureaucrats, politicians, and directors within each time period. There is a clear trend of increasing concentration from 2004 to 2018, showing that not only are new links being formed rapidly between the government and corporate sectors, but the degree centralities of the interlocking nodes are increasing as well, leading to an increase in the value of the indicator.

We further investigate the interlocking network to check if it is becoming denser itself. The clustering coefficient (CC) of the interlocking network does not increase over time (Figure 5), showing that the bridge edges are being formed between different pairs of entities. However, the CC of the 1-hop, 2-hop, and 3-hop neighborhood of the interlocking network extended only on the corporate side shows an increase over the years, indicative of increasing connectivity within the corporate network, either through subsidiary links or shared members in the BoDs of companies. A similar investigation of the neighborhood of the interlocking network extended on the government side however does not show any increase in the CC, validating that the increasing interlocks are happening due to increased concentration in the corporate network, and with more links being formed between bureaucrats and the corporate sector.

We take this further and go deeper into the corporate network to understand the dynamics there. We first consider only the network of company ownership formed by owner-subsidiary links, and call it the 'S' (subsidiary) network. We then also consider links between companies if they share one or more common directors, and call this the 'SP' (subsidiary-people) network. Figure 6 shows changes in the clustering coefficient and the size of the largest connected component (LCC) of the 'S' and 'SP' networks over time, for the 500 highest income public companies. The fact that the CC of the 'SP' network is 2x that of the 'S' network, and the LCC is more than 10x, is a strong indication of the importance of directors in increasing the density of the corporate network. These trends strongly point towards the formation of power structures by directors in the corporate network. Among these directors, we also find a significant presence of bureaucrats. In fact, among these directors of the 500 highest income companies, 2.27% (80 in total) are former bureaucrats who indirectly connect these companies to their former government departments (we see 2 and 3 hop connections like department—bureaucrat—company and department—bureaucrat—company—company). *Finance, commerce and industry, and textiles* are among the top most heavily interlocked departments connecting to some of the largest firms through bureaucrats, which can influence policy in specific direction. Other than the macro-level analysis presented above, the data can also be used to analyze micro-level patterns through which interlocks between specific entities are strengthened. Figure 7 shows some interesting anonymised network motifs, where relationships between a pair of individuals is strengthened over time through new connections. Such time-based patterns can be spotted in the data and flagged for further investigation.

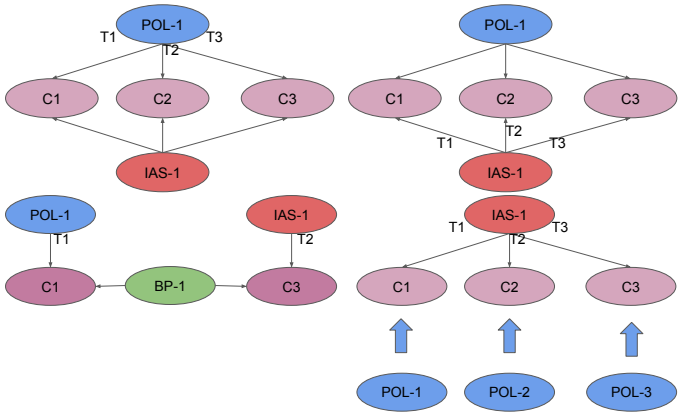
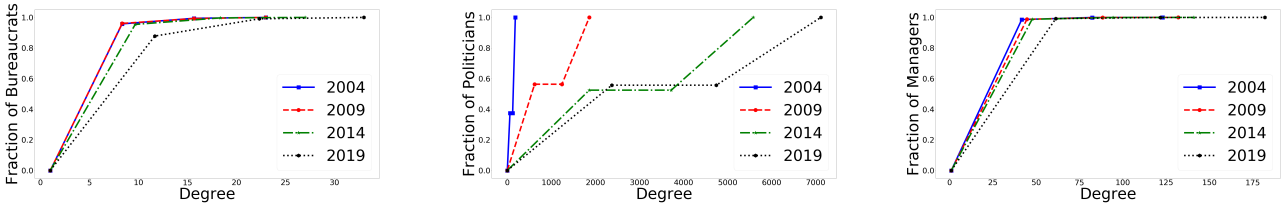


Figure 7: Different interlock motifs from top left, clockwise: (a) Politician joining firms connected to the same bureaucrat (b) Bureaucrat joining firms connected to the same politician (c) Politician and bureaucrat joining firms connected by a common business-person (BP) (d) Bureaucrat joining firms connected to important politicians through multiple hops. T1, T2, T3 show the timestamps of joining the firms.

Although in this paper, we have developed a country level indicator of corporate-government interlock, our methods can be applied within a specific policy area to understand the political economy in which it operates.

7 THE NEWS NUGGET APPLICATION

As discussed earlier, corporate-government interlocks sometimes manifest themselves in various outcomes, including the manipulation



(a) Bureaucrats (b) Politicians (c) Directors

Figure 4: CDFs of degree centralities of interlocking bureaucrats, companies, politicians, and directors

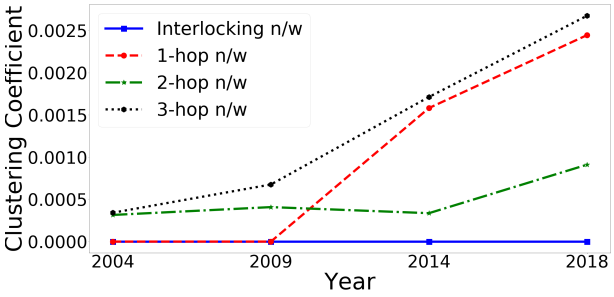


Figure 5: Change in clustering coefficient of the corporate-government network with time

of policies and scams. The goal of the *nugget application* is to make it easier for users to interpret specific policies or events, by knowing about the main actors involved in them. The application does this by extracting the mutual relationships between these actors and their links with other stakeholders in the ecosystem, and obtains news articles which might carry more information about these linkages. The application works as follows. It takes as input a query string about the event, consults the media database to find a set of relevant articles associated with the event, identifies the most frequently occurring entities in at least three news sources, then obtains a subgraph of interconnections between these entities by consulting the social network graph, and finally uses a simple ranking heuristic to find relevant articles which maximally cover the subgraph.

Subgraph extraction: Our goal is that given a set of target entities, we want to obtain a succinct subgraph of interconnections between these entities. To do this, we first offline pre-compute all paths of up to eight hops (the adjacency matrices used in pattern ranking has a maximum diameter of 8) between all pairs of entities, and rank these paths based on the average of the PageRank scores of the entities on the paths. Then, given the target set of entities for the subgraph extraction, we enumerate the top five paths between all pairs of these entities, and pick those paths which give the maximum coverage weighted by the entity scores of the entities. A limitation of this approach is that the resultant subgraph may not be connected, and we are improving our methods by building an optimization algorithm which can obtain a high density connected subgraph of the most important entities. As a case study, we consider an important event *2G Scam*, when many ruling politicians were accused of illegally allocating 2G cellular spectrums to telecom firms, and high ranking government officials and politicians were found to have worked in collusion with executives in telecom companies, to influence the auctions for frequency allocations in favor of these companies. The politicians were found to have profit making links to these companies through a web of other companies and family members. We obtained 11306 articles corresponding to the 2G scam from our media corpus, from which we obtained 20 most prominent entities based on their frequency of occurrence. The nugget app returned the interconnection subgraph between these entities as shown in figure 8.

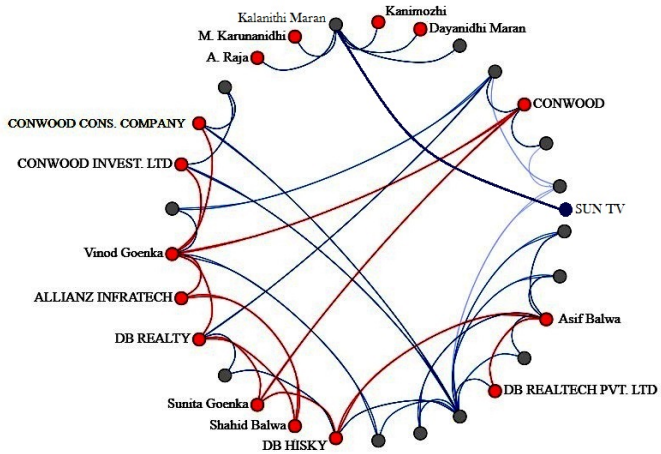
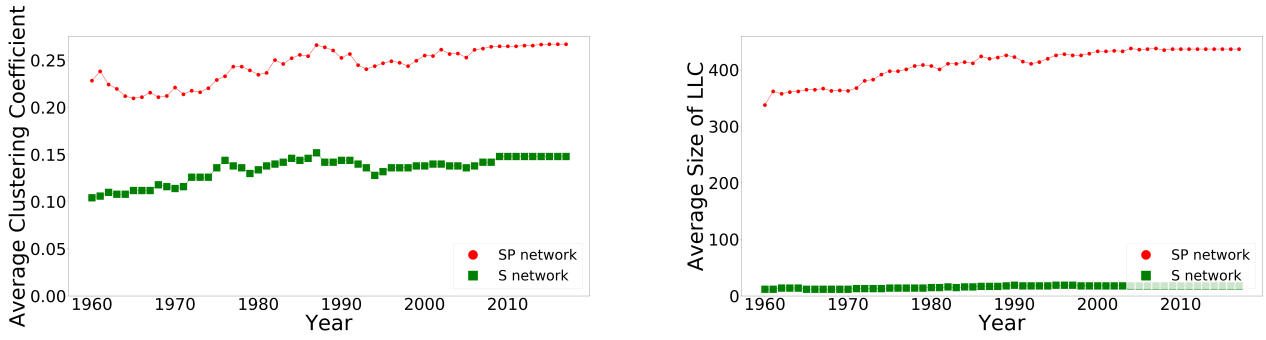


Figure 8: Subgraph returned by the nugget app for 2G scam: the entities in red are the important actors extracted by our frequency based retrieval measure. Edges and the blue nodes show the latent edges and nodes, extracted by the app from the social network.

In the figure, the red colored entities are the target 20 entities, and the blue entities are other important entities (with high scores) which interconnect these target 20 entities. It is interesting that even with this simple method of subgraph extraction, we are able to discover several entities and relationships which were not highlighted in any of the media articles published in the mainstream newspapers. These previously unknown connections (obtained from the social network



(a) Clustering coefficient

(b) Size of largest connected component

Figure 6: Plot of the clustering coefficient and largest connected component size of top 500 high income public firms over time

constructed from non-news sources) which enrich the context of the information obtained from news articles, is one of the main motivations for the *nugget application*.

Article ranking: Given a subgraph, the application also outputs a ranked list of news articles based on the importance of the entities mentioned in the article and the amount of subgraph coverage achieved by the article. The rank score for an article is computed as the product of two quantities: the fraction of the subgraph covered by the article, and the sum of entity scores of the entities covered. To evaluate the performance of our article ranking scheme, we a set of 50 interesting articles on the 2G scam covering most of the prominent entities were manually ranked by two researchers and compared with the ranking produced by our article ranking heuristic, which gave an NDCG score of 0.46. Although not very high, we are evaluating other methods for subgraph extraction and article ranking as well, and we feel that the discovery of new relationships could be a useful indicator for experimentation.

8 DISCUSSION

In this paper, we present a system that can be used to extract useful knowledge from web and media data to monitor the extent of corporate-government interlocks in India, and study its manifestations in the form of scams or policy issues. Our key findings from this analysis are as follows: (a) There is a monotonic increase in the extent of corporate-government interlocks over time, in the 2014-2018 period. (b) This increase occurred primarily due to the increase in degree centralities of the interlocking nodes. (c) Although the density of the interlocking network did not increase over time, the densities of the 1/2/3 hop neighborhood of the interlocking network increases significantly on the corporate side (d) Corporate interlocks created by company directors are the primary entities responsible for this increase, and not an expansion of company ownership networks through subsidiaries. We also developed a nugget application that aids users in studying scam events (from media data) through which such interlocks can exercise their influence.

So far, we have presented our findings based on explicit interlocks between different corporate and government entities. Implicit relationships between different institutions and communities too significantly shape policies, some unconsciously such as a growing similarity in educational and income backgrounds of politicians as more neo-liberal aligned thinking permeates the institutions, and some even consciously exacerbated by lobby groups and media to influence policies in specific industry sectors. In another work, we are also attempting to study these implicit interlocks by analyzing mass media and social media views expressed by different people about various events and topics, to discover an alignment of views. We hope this will help us answer important research questions with the aid of computation, such as whether changes in explicit and implicit interlocks between the government and corporate institutions coincides with outcomes of these interlocks in policy formulation. We feel this is an important dimension of citizenship that people in democratic functioning governance systems should be aware of, to be able to impose suitable checks in shaping the society.

9 CONCLUSION

There is a strong need to study corporate-government interlocks in a systematic manner. Our work contributes towards this by helping users track changes in the interlocks over time, and critically examine news articles about scams and policy discussions keeping in mind a view of the interconnections between entities involved in these events and topics. We have described in this paper a framework to build applications by combining information from news articles published in mainstream newspapers, and a knowledge base of important government and corporate entities along with their relationships. We have described several challenges we faced in putting this together, and algorithmic components we developed for data analysis and presentation. Our platform is able to successfully point out several interesting patterns of growth and transformation of interlocks which are worthy of further investigation by journalists and researchers, and help users get succinct views of complicated interconnections that shape scams and policy changes. Going forward, we are improving our computation methods and robustness of the platform, and aim to eventually release an easy-to-use publicly accessible website which users can browse to understand corporate-government interlocks, and possibly even contribute additional information in a crowd-sourced manner to enrich the dataset.

ACKNOWLEDGEMENT

We thank Gilles Verniers and Sudheendra Hangal from Ashoka University for having provided crucial insights and data which greatly assisted the research. We are also grateful to Shivam Rana from KPMG for his contribution to the work.

REFERENCES

[1] Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*. ACM, 85–94.

[2] Facundo Alvaredo, Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2017. World inequality report 2018. *The World Inequality Lab, http://wir2018. wid. world* (2017).

[3] Thomas M Begley, Naresh Khatri, and Eric WK Tsang. 2010. Networks and cronyism: A social exchange analysis. *Asia Pacific Journal of Management* 27, 2 (2010), 281–297.

[4] Marianne Bertrand, Francis Kramarz, Antoinette Schoar, and David Thesmar. 2006. Politicians, firms and the political business cycle: evidence from France. *Unpublished working paper. University of Chicago* (2006).

[5] Indrajit Bhattacharya and Lise Getoor. 2006. A Latent Dirichlet Model for Unsupervised Entity Resolution.. In *SDM*, Vol. 5. SIAM, 59.

[6] Mikhail Bilenko and Raymond J Mooney. 2003. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 39–48.

[7] William W Cohen and Jacob Richman. 2002. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 475–480.

[8] Peter Enderwick. 2005. What's bad about crony capitalism? *Asian Business & Management* 4, 2 (2005), 117–132.

[9] Mara Faccio. 2006. Politically connected firms. *The American economic review* 96, 1 (2006), 369–386.

[10] Raymond Fisman. 2001. Estimating the value of political connections. *American Economic Review* (2001), 1095–1102.

[11] Edward S Herman. 1988. Manufacturing consent: The political economy of the mass media (2002, Edward S. Herman and Noam Chomsky; with a new introduction by the authors.; Updated ed. of: Manufacturing consent. c1988.; Includes bibliographical references and index. ed.). (1988).

[12] The Hindu. 2014 (updated June, 2016). Coal scam: Chronology of events. (2014 (updated June, 2016)). http://www.thehindu.com/news/national/coal-scam-chronology-of-events/article6350481.ece

[13] Public Accountability Initiative. Accessed on Jan 2018. Littlesis. (Accessed on Jan 2018). https://littlesis.org/

[14] Lakshmi Iyer and Anandi Mani. 2012. Traveling agents: political change and bureaucratic turnover in India. *Review of Economics and Statistics* 94, 3 (2012), 723–739.

[15] Simon Johnson and Todd Mitton. 2003. Cronyism and capital controls: evidence from Malaysia. *Journal of financial economics* 67, 2 (2003), 351–382.

[16] Kalev Leetaru and Philip A Schrodt. 2013. GDELT: Global data on events, location, and tone. In *ISA Annual Convention*. Citeseer.

[17] Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2714–2721.

[18] Atif R Mian and Asim Ijaz Khwaja. 2004. Do lenders favor politically connected firms? Rent provision in an emerging financial market. *Rent Provision in an Emerging Financial Market (December 2004)* (2004).

[19] Charles Mills. 1956. Wright: The power elite. *New York* (1956).

[20] Thomas Piketty. 2014. Capital in the 21st Century. (2014).

[21] Thomson Reuters. Accessed on Jan 2018. Open Calais. http://www.opencalais.com/. (Accessed on Jan 2018).

[22] Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 623–632.

[23] Parag Singla and Pedro Domingos. 2006. Entity resolution with markov logic. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 572–582.

[24] John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization* 7, 2 (2008), 118–132.

[25] Joseph E Stiglitz. 2012. *The price of inequality: How today's divided society endangers our future*. WW Norton & Company.

[26] Sandip Sukhtankar. 2012. Sweetening the Deal? Political Connections and Sugar Mills in India. *American Economic Journal: Applied Economics* 4, 3 (2012), 43–63. http://www.jstor.org/stable/23269730

[27] Chris Taggart and Rob McKinnon. Accessed on Jan 2018. OpenCorporates: The largest open database of companies in the world. (Accessed on Jan 2018). https://opencorporates.com/

[28] Elasticsearch Developers' Team. Accessed on Jan 2018. Elasticsearch. (Accessed on Jan 2018). https://www.elastic.co/

[29] Neo4j Developers' Team. Accessed on Jan 2018. Neo4j. (Accessed on Jan 2018). https://neo4j.com/

[30] Chad Vicknair, Michael Macias, Zhendong Zhao, Xiaofei Nan, Yixin Chen, and Dawn Wilkins. 2010. A Comparison of a Graph Database and a Relational Database: A Data Provenance Perspective. In *Proceedings of the 48th Annual Southeast Regional Conference (ACM SE '10)*. ACM, New York, NY, USA, Article 42, 6 pages. https://doi.org/10.1145/1900008.1900067

[31] Wikipedia. Updated Jan 2018. 2G spectrum scam. (Updated Jan 2018). https://en.wikipedia.org/wiki/2G_spectrum_scam

[32] Wikipedia. Updated Jan 2018. Indian Administrative Service. (Updated Jan 2018). https://en.wikipedia.org/wiki/Indian_Administrative_Service

[33] Wikipedia. Updated May 2018. *2016 Indian banknote demonetization*. Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/2016_Indian_banknote_demonetisation

[34] Shyamlal Yadav. 2011 (updated Jan 2018). Bureaucrats: Life begins at 60. (2011 (updated Jan 2018)). http://indiatoday.intoday.in/story/life+begins+at+60/1/125878.html

[35] Chengqi Zhang and Shichao Zhang. 2002. *Association rule mining: models and algorithms*. Springer-Verlag.