Revolving Doors

A Needle in a Data Haystack – Final Project

Ella Neeman
ella.neeman@mail.huji.ac.il, ella

Gal Patel
gal.patel@mail.huji.ac.il, gal.patel

Dan Amir
dan.amir@mail.huji.ac.il, dan.amir

I. Problem Description

In a political context, ”capital-government relations” describes an erroneous and tendentious decision-making on behalf of government officials and institutions with the intention of obtaining personal or constituent benefits. This corrupt governmental conduct manifests in many ways, one of which is the ”Revolving Doors” phenomenon, describing the frequent two-way occupational shifts that are made between those who hold public offices and leading industry executives. A ”Revolving Doors” situation may cause regulation and policy-making to deviate from their original goals, as decision-makers are ever aware of the role they may be filling once they retire from their public positions. Even in the absence of an explicit give-and-take contract, the consideration of future career prospects distorts the regulator's objectivity and places them in a constant conflict-of-interest.

In this project our goal is to investigate this phenomenon in Israel. We seek to identify transitions of personnel from positions in government offices and other government institutions to companies linked by financial or other interests to those entities in the public sector. As a sub-goal, we aim to automatically identify such relations between the public and private sectors. Specifically, we would want to find pairs, or larger sets of entities, from both sectors, where the companies can benefit from the influence of the political organizations.

The idea of our project is based on a previous project initiated by ”Hasadna” (The Public Knowledge Workshop[1]) and has been frozen in the data collection stage. Therefore, we used some of Hasadna's resources and code base.

II. Data

This project required a complex integration of data from several different sources, some of it already collected by Hasadna and cleaned and processed by us, and some was collected from scratch.

A. Inetersts Map

In order to map the interests between entities in the Israeli economy and political institution we used two datasets:

- **Knesset Committee Meetings -** We used and improved code written by Hasadna to scrape protocols of Knesset committee meetings, extracting from them lists of guests while parsing the text. For each guest we extract its name and organizational affiliation. We obtained 254,324 entries of attendees in 17,282 meetings of the Knesset committees between the years 2004-2017. In order to tackle the challenge of finding the attendees names and descriptions in an almost unstructured text, we extended a parser written by Hasadna, which extracts full names and isolates them from the company name based on lexicons of Hebrew names, also using word order heuristics. We cleaned companies' names by truncating noisy prefixes that tend to appear in the protocols with them (like ”CEO” and others). We gathered that prefixes list by looking at a sample of the parsed records. We assume that entities which tend to appear at the same meetings will probably have some joint or conflicting interests.
- **Ynet Economy Articles -** We scraped Ynet economy channel homepage, and extracted the textual content of 29,408 articles which appeared on this page between the years 2004-2019. Since there is no archive available online which lists the articles, we used Wayback Machine platform to access older versions of this page. In order to identify mentions of organizations in those articles, we used Hebrew Named Entity Recognition (NER) tool from the python library polyglot. Since such tools tend to be more accurate for English text, we also experimented with extracting articles from the English version of Ynet and using Spacy NER extraction tool. We found the results in English to have similar quality while the amount of textual data from the English version of the site was significantly smaller, thus we used only the data from the Hebrew version.
  Although the predictions of the NER extraction tool were very noisy we found that filtering out terms with low discovery ratio - the ratio between the number of occurrences of the term discovered by the NER extractor and the total number of occurrences of the term in the corpus - results in cleaner data with less false-positive mistakes while keeping most of the correct predictions. We also eliminated any term which appeared less than 10 times in the corpus as the statistics for such terms will be inaccurate.

We decided to use those two data sources as we assumed they can well complement each other. Committees data is less noisy and includes information about relations which are not in the focus of media coverage (like volunteer

[1] https://www.hasadna.org.il/

organizations) while Ynet include more information about private companies.

B. Names Matching

One of the main obstacles with such data (names extracted from free text) is the large number of textual variants of terms which refer to the same entity. For example the terms: 'המועצה להשכלה גבוהה', 'מל"ג' and 'ות"ת' all refer to the same organization. Other variants are ones resulted form typos in the committee protocols and cases where the NER extractor captured only part of the entity name. Thus, our goal was to find such sets of strings which should be mapped to the same entity. In order to achieve that goal we combined several approaches, both general and content specific, for each data source:

- **String Matching -** In many cases the strings which refer to the same entity will be similar textually, thus we can use tools for fuzzy string matching to extract pairs which almost surely should be merged to one entity. We used the fuzzywuzzy open-source python library to obtain string similarity scores for pairs of entity names.
- **Common person names -** In the case of committee meetings, we expect that different names of the same entity will appear with common person names in the guests lists. This similarity can be measured as Jaccard similarity between the sets of attendees' names which appeared with each entity string. When used naively, this approach can yield many false-positive matches as there are names like משה כהן which are common in general and thus their appearance in conjunction with two entity strings isn't indicative. Thus, we first filter person names which appeared with at most five entities.
- **Wikipedia Search -** We used wikipedia's python API to search for pairs of terms which direct to the same wikipedia page, assuming that in most cases those will be terms that refer to the same entity. We observed that many times the entity will not appear as the first result in the wikipedia search. For example, searching for 'אסם' yields in the first suggestion 'ASAM' (Barn). Thus, relying on the fact that such mistakes are relatively rare, we collected the Wikipedia page categories for all the first results for terms in our Ynet data and used the most common categories as indicators of whether the wiki page refers to an organization (after manual filtering of some too general categories). Then, we choose the Wikipedia page for each term to be the first one in the top-5 results which belongs to one of the indicative categories. If no such result appears in the top-5, we take the first result regardless of its category. Since this method was too slow we used it only for the Ynet data as it included less terms than the Committees data.

Merging Procedure for Committees Guests:

We started with fuzzy string matching to first find "easy" matches and merged any pair with score higher than 88 out of 100. The score was chosen by randomly sampling pairs with scores in descending order and stopping when we start seeing false-positive matches. Since the number of potential pairs was too large to compute, we considered only pairs with at least one common person. In the second phase, since we already merged many sets of terms into single entity, we could now rely on the ”Common Person Names” score to merge pairs which are not similar textually, like ”General Electric” and ”GE”. We therefore picked all pairs with Jaccard score of at least 0.1 and at least 2 joint person names, and all pairs with score of at least 0.05 and at least 3 joint person names. We hypothesized that since the person names score become more accurate as more strings are merged (persons which were considered non-indicative as they appeared with too many entities, later become more indicative when those entities were merged into the same one), we could benefit from performing this step iteratively, but we found out that the merges didn't affect the Jaccard scores for most pairs. We then extracted more pairs by combining the person name score and the fuzzy matching with lower thresholds for both (0.01 for Jaccard and 70 for fuzzy matching). Finally, after a last merging refinement process which will be described later, we filtered all entities which appeared in less than 5 meetings.

Merging Procedure for Ynet data:

We extracted any pair with fuzzy string matching score higher than 83 and partial fuzzy matching score higher than 89. Partial matching refers to a variant of the same fuzzy matching score which takes for two strings the best match between sub-strings of length equals the minimal length of the two strings in the pair. Since the number of entities and pairs in the case of the Ynet data was much smaller, we filtered the results manually. We then used the Wikipedia search procedure described earlier to find pairs to merge without textual similarity.

Matches Correction:

Although the previously described merging procedure generates much cleaner and consistent data, any mistake in this stage could dramatically effect the results of our analysis later on. One way to look at this procedure is as extraction of connected components from a graph where each entity string is a node and edges are links between strings which resulted in a match in our procedure. We expect mistakes to be bottlenecks in our graph as they connect two sets representing two distinct entities which usually are densely connected. To identify merge sets with such bottlenecks we computed the second smallest eigenvalue of the laplacian matrix of their corresponding graph, also called the algebraic connectivity of the graph. We then visualized the graphs of merge sets with low algebraic connectivity to find wrong merges (see Fig. 1 for example of such merge set). We also generated automatic suggestion of edges to remove in each graph based on the maximal betweeness score.

Applying both the filtering and merging procedure for the Ynet and committees data resulted in a set of 2,561 unique entities.

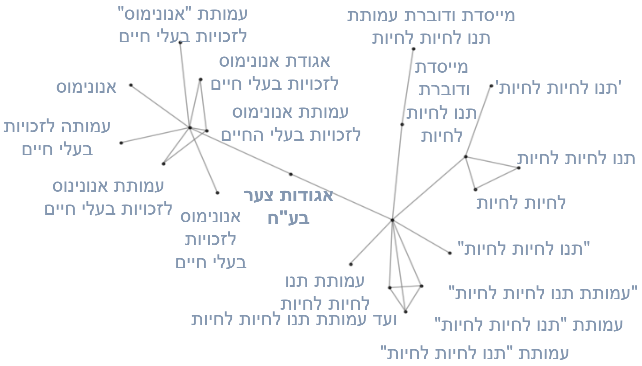


Figure 1: Example of a graph generated for a set of names with low algebraic connectivity (0.038). It can be seen that two entities were merged together linked through a node which described a general definition which includes both organizations. The suggested edge to break based on betweeness was אגודות צער בע”ח, עמותת תנו לחיות לחיות

C. Directors Appointments

In order to map transitions of personnel from the public sector to the private sector, we used Hasadna's code to scrape appointments reports from Maya - The Tel Aviv Stock Exchange Corporate Actions Systems. Maya provides a view of corporate events and income payments dates including reports about appointments of board of directors members and management staff in companies. We extracted data about 14,933 appointments between the years 2004-2019. Each data entry includes the name of the company and its ID number from the companies registrar as well as a list of previous jobs which can be used to detect the appointments we are interested in.

D. Entities Database

We merged several databases created by Hasadna to a unified entities database which includes: active association from the associations registrar, entities from the Government Companies Authority (GCA) public information, companies from the companies registrar website, list of cooperative societies from the Ministry of Economy website and others. The unified database includes 471,404 entities where each entry includes the entity official name, unique ID (equivalent to the ID in the companies registrar when such exists) and classification of the entities to different types (see Fig. 2). We then matched each entity from our Interests Map data (Ynet + committees) to the database by choosing the one with the highest fuzzy matching score, using the set of names which was merged with it, if such available. More precisely, if x denotes a name of entity from the interests map data, M_x denotes its merge set, f denotes the fuzzy matching function and $\mathcal{D}$ denotes the set of names in the entities database. We choose the match by the following rule:

$$m(x) = \arg\max_{y \in \mathcal{D}} \max_{x' \in M_x} f(x', y)$$

Company		
company	חברה פרטית	provident_fund
foreign_company	חברה ציבורית	professional_association
association	חברת חו”ל	health_service
ottoman-association	שותפות מוגבלת	university
חברה פרטית מחוייבת במאזן	private	cooperative
Government		**Other**
Government_office	drainage_authority	municipal_precinct
law_mandated_organization	conurbation	house_committee
local_planning_committee	municipality	foreign_representative
religious_court_sacred_property	religion_service	municipal_parties

Figure 2: mapping between organization types found in the merged entities database to our three super-types: government, company and other.

Since calculating fuzzy matching between every pair of names from the interests map data and entities database appeared to be too time consuming, we first filtered, for each name, only the entities in the database which contains the name as a substring.

Matching the entities in out data to this general entities database allows us to focus on finding relations between pairs of entities we are interested in (inter-sector and not intra-sector) and match appointments from Maya with entities we found in the interests map based on the companies registrar ID.

III. Methods and Results

As described earlier, the first goal of our project is to map the relations between entities in the private and public sector. Those relations can then be used to find employees transitions which may indicate political corruption. We considered here two main approaches for analyzing the data - analyzing the relations as a graph and extraction of association rules.

A. Graph structure

Based on the data from Ynet articles and Committee meetings, we can calculate a similarity metric between pairs of entities and analyze the graph constructed from weighted edges based on this metric. We considered two options for similarity metric:

- **Jaccard Similarity** - We can represent each entity as a set of the articles/meetings in which it appeared and then calculate Jaccard similarity for pairs of entities based on their corresponding sets. We divided the scores by the maximal value over all pairs to normalized the range to 0-1 (so it can also represent adjacency on a graph). Since we found the distribution of Jaccard values

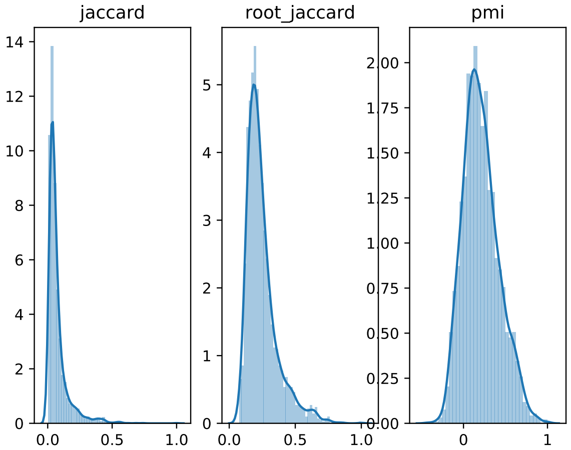


Figure 3: Histograms comparisons for different similarity scores - Jaccard, the squared root of Jaccard and PMI. All scores are first normalized to have a maximal value of 1.

to be very left skewed and centered near 0 (see Fig. 3), we used the squared root of it as adjacency measure.

- **Mutual Information** - Another approach we considered is treating the appearance of entity or pair of entities in a set as probabilistic event and calculate the pointwise mutual information between the indicator variables. This results in the formula:

  $$PMI(x_1, x_2) = \log\left(\frac{Pr_{y \sim Y}(x_1 \in y \wedge x_2 \in y)}{Pr_{y \sim Y}(x_1 \in y) \cdot Pr_{y \sim Y}(x_2 \in y)}\right)$$

  Where x_1, x_2 are entities, and Y is the set of possible articles/meetings. If we estimate those probabilities with our empirical data we get after some rearranging of the equations the following term which is also equal to the log of the lift of the items-set (x_1, x_2):

  $$s(x_1, x_2) = \log\left(lift(x_1, x_2)\right)$$

  Where:

  $$lift(x_1, x_2) = \frac{|\{y \in Y : x_1, x_2 \in y\}| \cdot |Y|}{|\{y \in Y : x_1 \in y\}| \cdot |\{y \in Y : x_2 \in y\}|}$$

  We clip negative PMI values as anti-correlation of entities appearance doesn't have a clear meaning in our case.

In order to choose a similarity metric, we extracted the pairs with highest absolute difference in rank based on each of the metrics as well as the set asymmetric differences between their two top-50 results. We also compared qualitatively clustering results based on both metrics. Since we couldn't find significant difference in results quality, we chose to use the PMI as we thought it is more grounded from theoretical point of view and it's meaning relates directly to the quantity we want to measure.

Another decision we had to take when analyzing this data as a network graph is how to combine the two sources of data - Ynet articles and committees. We first confirmed that our hypothesis that those data sources are not redundant by examining the histogram of organization types in both datasets (see Fig. 4). Based on the observation that, in our case, the absence of similarity (or low similarity) doesn't imply absence of interest or relation (since it could be caused for example by limitations of data coverage), we decided to take for each pair the maximal value between the similarity scores obtained from each dataset. We also filtered any pair of entities with number of low joint occurrence count (0.0002 of articles in Ynet, and 0.0008 of meetings in the committees data) as for those pairs the estimation of PMI will be inaccurate. In order to allow use of pairs with lower count, we also tried weighting the estimate based on this joint count, using the following formula:

$$s_{weighted}(x_1, x_2) = lift(x_1, x_2) \cdot \frac{c(x_1, x_2) - m}{c(x_1, x_2)} + \frac{m}{c(x_1, x_2)}$$

Where $c(x_1, x_2)$ is the joint count of the pair and m is the minimal count used for filtering pairs. Intuitively this term is a weighted average between the observed estimate value and a null hypothesis of no relation (which equivalent to a lift of 1) where the count is used to weight between the observation and the no-relation ”prior”. We found this fix to have minor effect on the results, thus we used the simpler score and the minimal count threshold described earlier.

We used spectral clustering (with $k = 20$ - based on observed gap in the eigenvalues of the laplacian matrix) to find clusters of entities with similar interests. In order to examine the results we visualized the graph and clusters using the holoviews library (see Fig. 5 , or provided html files for the interactive version). We found most of the clusters to be meaningful though some recurring patterns of mistakes appeared - large government offices tended to be merged together to the same cluster even if their area of interests is only partially connected, and some clusters include two or more areas of interests linked by few unifying entities. We tried solving the first problem by focusing only on a bidirectional graph between entities from the private sector on one side and entities from the public sector on the other, we then used spectral bi-clustering (also called co-clustering) to find blocks in the asymmetric adjacency matrix (where each axis includes entities from one sector). We compared the results to the standard spectral clustering but found little to no difference. In order to solve the latter issue, we tried increasing the number of clusters which resulted in noisy predictions and arbitrary partitions of meaningful clusters from the previous results.

B. *Association Rules*

As another approach to map the connections and interests between sectors we employed the use of association rules. We treated our data of Ynet articles and committees meetings as transactions data: articles and meetings act as baskets, while entities are items. Using the Apriori algorithm (from the Apyori python's library), we extracted association rules

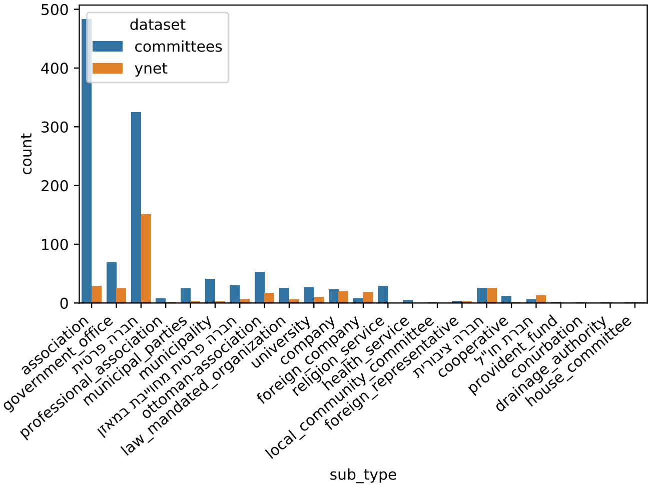


Figure 4: Histograms of entities sub-types in Ynet data and committees data. It can be observed that the ratio between associations and companies differs drastically between both datasets.

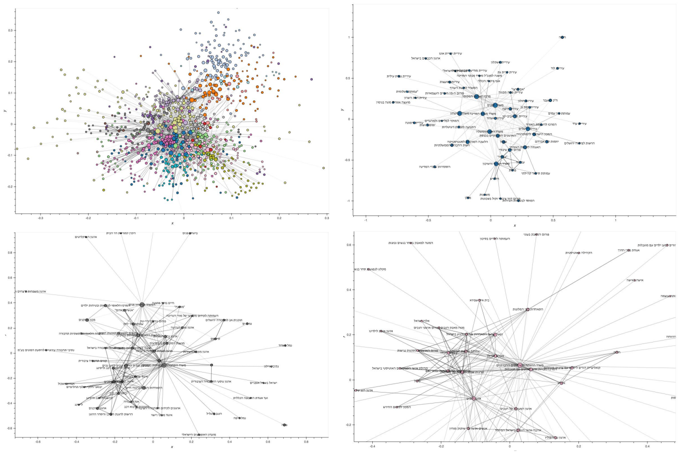


Figure 5: Graph visualization of the interests map - layout of all the entities and some chosen clusters. Layout generated by Fruchterman and Reingold algorithm, nodes are colored based on spectral clustering and their size determined according to the number of appearances in articles and meetings. Zoom in to see labels of entities in the specific clusters.

which consists of single-item-sets of the form $\langle Company \rangle \rightarrow \langle Governmentbody \rangle$. We chose to focus on this direction based on the assumption that government entities tend to be involved in many domains while the connection to the private company will usually be restricted to a narrow interest. Our choice of evaluation is interest:

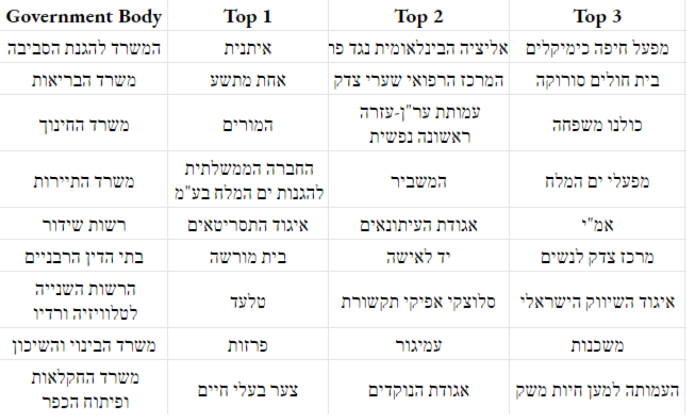
$$interest(x \rightarrow y) = confidence(x \rightarrow y) - support(y)$$

. We filtered the rules by:

$$support(x, y) \geq 0.0002$$

$$interest(x \rightarrow y) \geq 0.25$$

This way, we were able to map the interests of the private sector in different government bodies, with 2,999 association rules based on the committees meetings data and extra 9 based on the Ynet articles dataset (see Fig. 6 for sample from our results). We evaluated our results by sampling 30 random association rules and manually checking their correctness. By this measure, we got accuracy of 80%.



Government Body	Top 1	Top 2	Top 3
המשרד להגנת הסביבה	איתנית	אליציה הבינלאומית נגד פר	מפעל חיפה כימיקלים
משרד הבריאות	אחת מתשע	המרכז הרפואי שערי צדק	בית חולים סורוקה
משרד החינוך	המורים	עמותת ער"ן-עזרה ראשונה נפשית	כולנו משפחה
משרד התיירות	החברה הממשלתית להגנות ים המלח בע"מ	המשביר	מפעלי ים המלח
רשות שידור	איגוד התסריטאים	אגודת העיתונאים	אמ"י
בתי הדין הרבניים	בית מורשה	יד לאישה	מרכז צדק לנשים
הרשות השנייה לטלוויזיה ורדיו	טלעד	סלוצקי אפיקי תקשורת	איגוד השיווק הישראלי
משרד הבינוי והשיכון	פרזות	עמיגור	משכנות
משרד החקלאות ופיתוח הכפר	צער בעלי חיים	אגודת הנוקדים	העמותה למען חיות משק

Figure 6: A sample of association rules. Each row represents a government body and the companies that form a rule of the form $\langle Company \rangle \rightarrow \langle Governmentbody \rangle$ with the highest interest measure.

C. Appointment extraction

In order to merge all the insights we gathered, we split each appointment from Maya's data into a pair of new company and prior organization. We mapped them to our entities database, filtered to include only prior organizations in the public sector and then checked if they appear in the Interests Map extracted from Ynet articles and committees meetings, and kept only the Revolving Doors records.

IV. Results

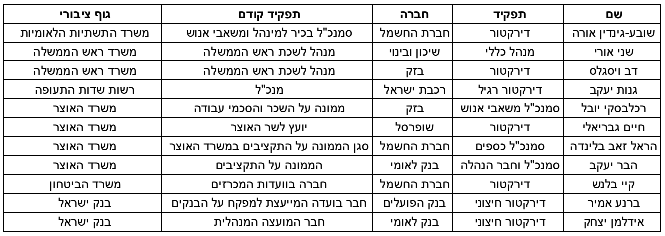
To measure the success of our project, we defined three evaluation senses:

- **Precision** - We collected some examples of occupational shifts that received media attention or published in an academic research framework.[2] We didn't expect to find lots of matches since this approach is very anecdotal, and indeed, most of the examples didn't appear in our data at all. One reason was because they weren't on Maya's appointments reports. For example:

[2] Roy Shapira, Revolving Doors and the Ineffectiveness of Cooling-Off Periods, Stigler Center, University of Chicago Booth School of Business; Interdisciplinary Center (IDC), 2019

גיא רוטקופף was מנכ"ל משרד המשפטים and his decision to become a consultant of חברת אבנר חיפושי נפט וגז מקבוצת יצחק תשובה isn’t documented on Maya’s reports. One Revolving Door we missed was the appointment of יהושע שוקי שי to בז"ן ־ בתי זיקוק לנפט after serving as a יועץ בכיר למנכ"ל משרד האוצר . We do hold the relevant entities in our data, and we succeed in passing this appointment to the interests check, but we fail in two terms: first of all, because of an unsuccessful string matching in the merging step, we can’t find the company by its id. ”בתי זיקוק לנפט” is mapped to ”צרכנית עובדי בתי הזיקוק לנפט בחיפה” , which is clearly wrong. Secondly, we determine that בז"ן has a clear interest only in המשרד להגנת הסביבה and משרד התשתיות הלאומיות .

- **Recall** - We planned to evaluate our results by taking a random sample of the output and checking that it ”makes sense”. Eventually, we got a small amount of records that represent problematic shifts (see Fig. 7), so we could evaluate them all. We found that all of the extracted cases of Revolving Doors were correct.
- **Novelty** - We wanted our results to contribute new insights about the Israeli politics, and to present unexpected relation between the public and private sector. In the interests mapping exploration we did find some non-trivial relations, for example, an association rule for interest between בית חולים תל השומר and רשות מקרקעי ישראל taught us about a plan to build 3,500 apartments in the hospital area[3], and revealed the hospital’s interest to avoid this action. Unfortunately, the poorness in the number of Revolving Doors records didn’t allow us to find trends in the data.



שם	תפקיד	חברה	תפקיד קודם	גוף ציבורי
שובע-גימדין אורה	דירקטור	חברת החשמל	סמנכ"ל בכיר למינהל ומשאבי אנוש	משרד התשתיות הלאומיות
שני אורי	מנהל כללי	שיכון ובינוי	מנהל לשכת ראש הממשלה	משרד ראש הממשלה
דב ויסגלס	דירקטור	בזק	מנהל לשכת ראש הממשלה	משרד ראש הממשלה
גנות יעקב	דירקטור רגיל	רכבת ישראל	מנכ"ל	רשות שדות התעופה
רכלבסקי יובל	סמנכ"ל משאבי אנוש	בזק	ממונה על השכר והסכמי עבודה	משרד האוצר
חיים גבריאלי	דירקטור	שופרסל	יועץ לשר האוצר	משרד האוצר
הראל זאב בלינדה	סמנכ"ל כספים	חברת החשמל	סגן הממונה על התקציבים במשרד האוצר	משרד האוצר
הבר יעקב	סמנכ"ל וחבר הנהלה	בנק לאומי	הממונה על התקציבים	משרד האוצר
קיי בלנש	דירקטור	חברת החשמל	חברה בוועדות המכרזים	משרד הביטחון
ברנע אמיר	דירקטור חיצוני	בנק הפועלים	חבר בועדה המייעצת למפקח על הבנקים	בנק ישראל
אידלמן יצחק	דירקטור חיצוני	בנק לאומי	חבר המועצה המנהלית	בנק ישראל

Figure 7: Appointment Extraction Results

V. Conclusions and Future Work

In this work we aimed to map both the links between entities in Israel’s politics and economy and the phenomenon of Revolving Doors - the transition of employees and directors between entities in the private and public sectors which are connected by interests. Extracting such information required collection, synthesis and construction of structured data about the entities and their relations. Given that the collected data was mainly in Hebrew and the absent of strong NLP tools in Hebrew, posed more challenges. We think this data can be further refined by developing NLP tools in Hebrew. One such tool can be an organization string matching algorithm which takes into account how indicative a words or sub-strings is, in the similarity calculation. For example matching of the word בע"מ between two strings should have low impact on the fuzzy matching score. Another way to extend and improve the entities matching between the different data sets would be to use more online resources, in a similar fashion to how we used Wikipedia.

When considering our suggested method performance, we find some mixed results. On one hand, we were able to identify particular events of Revolving Doors from a large appointments dataset (a real data in a haystack!) without false-positive mistakes. On the other hand, we believe that many more such events happened in the time frame analyzed in our data. The low number of discovered events is probably both due to inaccuracies in the entities mapping and lack of information regarding appointments in private companies and of lower rank employees. Two possible sources for such information that could be used in future work are: Linkedin user’s jobs history - an option we considered during the project planning but abandoned due to technical difficulties, and protocols of ועדת ההיתרים which is responsible for approving changes in cool-off periods for employees in the public sector.

We believe that further analysis of the interests map data can yield valuable insights both directly with respect to this subject and also in broader context with respect to politics and economics trends in Israel. One interesting approach could be to analyze those relations as dynamical system over time and finding changes in the map. On the algorithmic side, our analysis can be extended by unifying both our pairs based approach and graph structure approach and trying to exploit the higher order relations encoded in the graph (indirect connections trough paths) to compute more accurate pairwise similarity scores.

VI. Data Access

All data discussed in this paper can be access via Google Drive, here.

[3] Court gives green light to 3,500 homes in Tel Hashomer, Globes, Arik Mirovsky, 27 Feb, 2019, https://en.globes.co.il/en/article-court-gives-green-light-to-3500-homes-in-tel-hashomer-1001276046