UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Corruption Risk Profiling With Open Data: Application Of Graph Theory

Author:
Iryna POPOVYCH

Supervisor:
Liubomyr BREGMAN

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2021

Declaration of Authorship

I, Iryna POPOVYCH, declare that this thesis titled, "Corruption Risk Profiling With Open Data: Application Of Graph Theory " and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"We cannot understand our humanity just by studying individuals."

Nicholas A. Christakis

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Corruption Risk Profiling With Open Data: Application Of Graph Theory

by Iryna POPOVYCH

Abstract

One of the primary obstacles to the effective prevention of corruption is the lack of knowledge and understanding of the circumstances under which corruption arises and evolves. In this work, we develop an approach for open data processing aiming to build network models of public officials in Ukraine and identify relations between them as an indicator of corruption risk. The primary data source for this research is the Unified State Register of Declarations of Persons Authorized to Perform Functions of the State or Local Self-Government. We propose to single out the facts of social and economic relations between the subjects of the declaration (public officials), and on this basis, to model the networks using principles of social network analysis and graph theory. Our results show that the social networks of Ukrainian public officials can be modeled by defining relationships between them via company co-ownership or common participation in an organization. We explore the resulting networks with modularity coefficients, review communities within networks, and propose a framework for corruption risk profiling with the usage of network characteristics.

Keywords: *open data, graph theory, social network analysis, corruption, Ukraine.*

Acknowledgements

I am infinitely grateful to my mother, father, brother, my boyfriend, and all my family for their kind support and inspiration I received from them that helped me come to the finish line of my Bachelor studies and this thesis in a way I can be proud of it.

I am also thankful to my supervisor Liubomyr Bregman for his constant support, all the funny stories and experiences we had, and, of course, his fresh ideas for the research and valuable feedback.

Big thanks to Yulia Kleban, Head of IT and Business Analytics program, for all her patience, care, and support of the students during this four-year journey.

Many thanks to Ms. Nataliya Anon and Mr. Adrian Slywotzky for providing scholarships and the ability to study at UCU.

Finally, I want to thank all my groupmates, teachers, faculty, and university for an unforgettable experience of witnessing, serving, and communicating together.

And, of course, I want to say thanks to my dog, Dzyga, for all the love and mental support from her during the hard weeks...

Contents

Declaration of Authorship i

Abstract iii

Acknowledgements iv

1 Introduction 1
- 1.1 Overview of Open Data Principles . . . 3
- 1.2 Overview of Ukrainian electronic asset declarations system . . . 3

2 Discussion of related works and case studies 5
- 2.1 Open Government Data Research . . . 5
- 2.2 Social Network Analysis with the use of Open Data . . . 6

3 USRD Data overview and the approach to data modeling 8
- 3.1 Declarations API . . . 8
- 3.2 Data modeling . . . 9

4 Social networks of Ukrainian public officials and their characteristics 12
- 4.1 Topological characteristics of network graphs . . . 13
  - 4.1.1 Declarant-company and declarant-organization graphs . . . 13
  - 4.1.2 Declarant-company-organization graphs . . . 14
- 4.2 Reduction to declarant-declarant graphs . . . 15

5 Application of networks for corruption risk profiling 17
- 5.1 Important features of declarant-declarant graphs . . . 17
- 5.2 Possible ways of corruption risk profiling using network graphs . . . 19

6 Conclusions 20
- 6.1 Research summary . . . 20
- 6.2 Future work . . . 21

Bibliography 29

List of Figures

1.1 Ukraine CPI Score By Year (Transparency International, 2021) 2

2.1 OGD utilization framework (Safarov, Meijer, and Grimmelikhuijsen, 2017) 6

3.1 Declarations API simple JSON response 9
3.2 A sample snapshot of the final data set of 'company - declarant' pairs. Some attributes (columns) are hidden in this picture 11

4.1 The three types of nodes and the two types of relationships we define in the network graph 12
4.2 Community Sizes Distributions for 'declarant-company' and 'declarant-organization' graphs of Parliament and Prosecutors. 14
4.3 The indirect connection between declarants, via company and organization. 14
4.4 A snippet of Python code for graph simplification to a person-person graph. `people_nodes_graph` is a list of all person nodes, `parent_graph` is a composed declarant-company-organization graph. 15
4.5 The person-to-person relationship graph, colored by location of the declarant 16

5.1 A snapshot of the nodes with high betweenness centrality. The size of the node indicates its betweenness centrality 17
5.2 A declarant-to-declarant network graph of people from four different office types. 18

List of Tables

3.1 Number of unique declarations and their share in the final data set (annual 2019 NACP declarations) by type of relationship the declarant has with companies or organizations 10

4.1 Main topological characteristics of the declarant-company and declarant-organization graphs 13

List of Abbreviations

SN	**S**ocial **N**etwork
SNA	**S**ocial **N**etwork **A**nalysis
OGD	**O**pen **G**overnment **D**ata
USRD	**U**nified **S**tate **R**egister of **D**eclarations (of persons authorized to perform functions of the state or local self-government)
USRD	**U**nified **S**tate **R**egister of **D**eclarations
NACP	The **N**ational **A**gency on **C**orruption **P**revention
JSON	**J**ava**S**cript **O**bject **N**otation
EDRPOU	Ukrainian state registry legal entity identifier
NGO	**N**on-**G**overnmental **O**rganization
ID	**I**dentifier

For all those who value the truth…

Chapter 1

Introduction

Corruption is perhaps the dominant limitation for economic and social welfare. Since *secrecy* is a crucial factor for corruption, we can anticipate corruption with *openness* and transparency. Open data standards have been introduced in many countries to enable ongoing monitoring of economic and political processes. Dozens of countries publish information on property registries, transportation, public procurement, and many more. However, not all of this data is set up for automatic monitoring, let alone analysis. As a result, large volumes of published data do not bring many benefits to society. We understand that a significant contribution to the development of this area today can be the processing of large scale open data and its analysis.

Corruption is widely discussed in scientific publications. Researchers aimed to explore the causes and consequences of this issue (Mauro, 1995) (Tanzi, 1998). The global research confirms that this is a complex problem that can be approached from different perspectives. There were studies of corruption in specific sectors, like police (Punch, 2000), healthcare (Vian, 2008), or higher education (Rumyantseva, 2005). Also, corruption can be viewed on a country level, so there are studies focused on learning from the experience of particular countries like Singapore (Quah, 2001), or China (Wederman, 2004).

In most of these works, the researchers were focused on case studies and social aspects of corruption, using descriptive approaches or small data samples from questionnaires, which allowed them to suggest theoretical strategic solutions. However, as more governments digitize their operations and provide public access to information, there are opportunities for empirical research: using data to apply quantitative methods.

Ukraine is a country with a high level of corruption: in the 2020 ranking of Transparency International (Transparency International, 2021), Ukraine is down the list, in 117th place, with its Corruption Perception Index (CPI) being 33 out of 100. Even though we see a slightly positive trend for improvement over the last years, the pace is languid (see Figure 1.1).

Studies prove that corruption is often a part of relationships within political elites (Johnston, 1999). Also, several papers suggest that elite communities in Ukraine have strong power, and there are risks for corruption in elites networks (Kostiuchenko, 2012) (Kudelia, 2016).

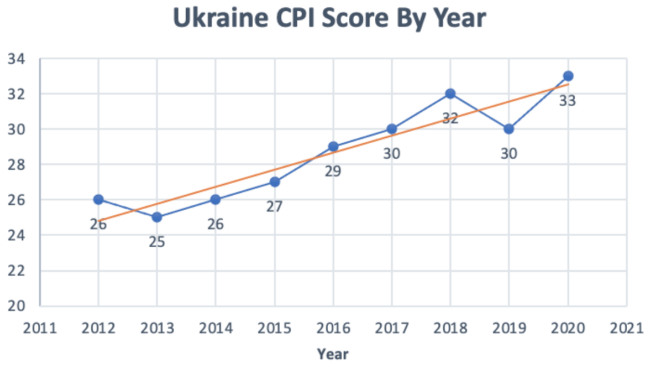


FIGURE 1.1: Ukraine CPI Score By Year
(Transparency International, 2021)

Another fact that is important to mention is that Ukraine has been working a lot towards open governance and publishing governmental data over the past ten years (The Open Data Institute, 2018). Many projects in Ukraine helped reduce corruption in various areas of the economy; for example, Prozorro (Prozorro, 2021), the public procurement system launched in 2015. The Unified State Register of Declarations of Persons Authorized to Perform Functions of the State or Local Self-Government (USRD) designed to monitor public officials' assets was digitized. We see that large data volumes are published, and, by this time, a significant portion of them is published in a machine-readable format. Still, very few data sets are constantly being processed and analyzed.

Given the above, we decided to do empirical research and discover the opportunity to use Ukrainian open data. This study is focused on searching for links between Ukrainian public servants using open data, modeling network graphs based on the links found, and analyzing the resulting networks. We have chosen to use USRD data for this work. The motivation behind this is that we have seen researches on public procurement data and corruption, but there is little or no research around USRD data.

The work in this paper is organized as follows: Chapter 2 is about the research that was done before us that is related to our topic; in Chapter 3, we describe the approach for data collection and processing and provide an overview of the data we used. Chapter 4 tells about our approaches to network modeling with this data; also, it gives the main characteristics of the networks we modeled. Chapter 5 describes the analysis of the networks and discusses possible applications of networks for corruption risk profiling. Chapter 6 provides the research summary and suggestions for future work.

1.1 Overview of Open Data Principles

Open data has gained much attention in recent years. According to the definition from the Open Data Handbook (Open Data Handbook, 2021),

> Open data is data that can be freely used, re-used, and redistributed by anyone - subject only, at most, to the requirement to attribute and share-alike.

The eight open data principles, defined back in 2007 (OpenGovData, 2007), remain substantive. According to them, open data should be:

- Complete
- Primary
- Timely
- Accessible
- Machine processable
- Non-discriminatory
- Non-proprietary
- License-free

The general idea behind this concept is that open access to information drives social transformations and economic development. The main advantage we see in using open data for our research is that it allows reproducibility and broad access to information. The opportunity to freely share the data sets and the insights from our work with others is a crucial factor for accelerating insightful discussions.

1.2 Overview of Ukrainian electronic asset declarations system

We use data from USRD for this research. The Register includes all the digital declarations of Ukrainian public servants from 2015 till now. The primary purpose of introducing asset declarations reform (2015) was to bring transparency and prevent corrupt practices by government officials. The National Agency on Corruption Prevention of Ukraine (NACP, 2021) is a central executive body responsible for developing anti-corruption policy and prevention of corruption. As a part of its operations, this agency owns and maintains USRD, where public officials submit their asset declarations.

Officials are obliged to declare all of their assets inside and outside Ukraine and all of their relatives' officially registered assets. The assets may include cash and

money in their bank accounts, precious gifts they received, real estate, transport, shares in companies, non-property rights, and more. All submitted files are publicly available.

Many public servants are supposed to fill their declarations. In general. We can divide them into two groups:

- Top state officials
- Local governments officials

The first group includes The President of Ukraine, The Prime Minister of Ukraine, Members of the Cabinet of Ministers of Ukraine, and other top officials. The local government group includes all the people who work in governmental structures, more than 300000 public servants.

Chapter 2

Discussion of related works and case studies

2.1 Open Government Data Research

The general movement towards transparency and open government has resulted in numerous national open data portals and infrastructures that provide Open Government Data (OGD) access. The evolution of these portals changes how both citizens and researchers use the data. Over the last years, many social innovators and scientists have started exploring OGD. Even though this area has been proliferating, most of the discussion about ODG happens outside the traditional research exchange. Instead, there are ongoing dialogues about OGD in various research institutions, non-government organizations, media platforms, and other foundations that often use formats different from traditional journal articles. Dozens of information about applications of OGD are represented in case studies, organization reports, or even blog posts, making it hard to provide a systematic review of all the associated research.

A work of fundamental importance in OGD application development is "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives" by Barbara Ubaldi (Ubaldi, 2013). It summarizes all the opportunities that OGD initiatives and analytics may offer policymakers and captures the challenges that could arise. Also, the work focuses on a transparent methodology for measuring the impact of open data initiatives. It suggests developing a standard set of metrics to assess impact in various countries consistently.

The introduction of this concept of impact measurements is essential because most of the previous research was focused solely on 'opening' the data and delivering it to the end-users, such as civil society. Insufficient attention was given to data quality, usability, and influence in the early passion for data openness. However, after significant data volumes have been published in recent years, the main research focus shifted to the practical side - discovering the usage of open data (and the lack of this usage).

One of the recent systematic analyses of practical usage of open data was presented in "Utilization of open government data: A systematic literature review of types, conditions, effects, and users" (Safarov, Meijer, and Grimmelikhuijsen, 2017).

Authors reviewed more than a hundred recent academic papers related to OGD utilization. They showed that most researchers focus on theoretical studies and build assumptions about using open data rather than empirically utilizing it. Based on the hypothesized relationships between open data users, the authors developed a multi-dimensional framework of OGD utilization (see Figure 2.1). The framework consists of four generic categories (users, effects, types, and conditions) and shows the plenitude of relations between these four categories; it also highlights that there is limited observed knowledge on the relations to the effects of OGD.

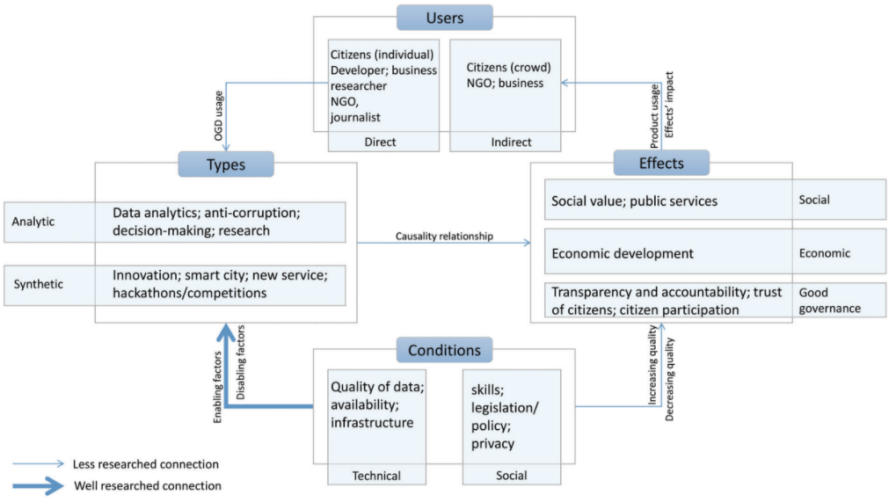


FIGURE 2.1: OGD utilization framework
(Safarov, Meijer, and Grimmelikhuijsen, 2017)

Many empirical pieces of research arose in the last years using OGD in various areas of social and economic participation. For example, a recent study on public procurement confirms that procurement officials are more likely to award treated contracts through open bidding after the EU open data initiative was introduced (Duguay, Rauter, and Samuels, 2020).

2.2 Social Network Analysis with the use of Open Data

Another emerging direction of studies is the analysis of networks built with open data. Can and Alatas (Can and Alatas, 2019) provide an overview of the variety of SN problems that are currently popular and present a comprehensive source of the studies performed in this area. Many researchers work with social graphs modeled based on data from online social networks like Twitter or Facebook. Backstrom and Kleinberg (Backstrom and Kleinberg, 2013) address the problem of recognizing a romantic partner of a person given the network structure of all the connections among a person's friends alone. After introducing the 'dispersion' - a new characteristic of a social network of friends, they reach the goal of their research with high accuracy.

Finally, the studies arise that combine OGD and SN modeling. Most of such works we observed were dealing with public procurement and contracting data.

"Using social network analysis in open contracting data to detect corruption and collusion risks" by D.H. Murillo (Herrera Murillo, 2019) uses a graph data model to represent the tendering behavior of a procurement network. Based on the network model, the author defines and describes actors and communities with unusual activity and implements a machine learning workflow from connected features to predict if a contract has been corrupted or not. The research also discusses limitations of the developed network approach for corruption detection: the lack of labeled data to construct the predictive models and the fact that it identified possible acts of corruption and collusion after consummated (the developed approach is reactive, not proactive).

Chapter 3

USRD Data overview and the approach to data modeling

3.1 Declarations API

This research uses the data of public figures' declarations of assets. There are two different sources of this information: NACP's official website (NACP, 2021), and an open data project, Declarations (Declarations.com.ua, 2021), built on top of NACP data. NACP is responsible for maintaining the register of declarations (USRD), and recent declarations are in open access on their website. Declarations project provides the largest database of declarations of public officials in Ukraine that includes all the e-declarations from NACP and manually digitized and proofread paper declarations (older ones). The platform provides open API access to all available JSON machine-readable data. Additionally, the platform provides functionality of built-in UI for searching and a built-in analytics module. In this paper, we decided to choose declarations.com.ua API as it provides large dataset, stable API, better documentation and search functionality which enabled more efficient analysis.

The API service allowed us to obtain declarations filtered by multiple parameters, such as region, body type (of the governmental body they are a part of), declaration year, location. It also allows a user to use any search query and apply full-text search across all declarations contents. Below there is an example of a simple request and the response from Declarations API. The query searches for all declarations from 2019 where declarants mentioned the word 'dogecoin' in any part of the declaration:

```
https://declarations.com.ua/search? \
      q=dogecoin&deepsearch=on&declaration_year=2019&format=opendata
```

This request returns a JSON (Figure 3.1) with information about the query, number of results, and found declarations in the `results.object_list` section. Description of the API response and the declaration structure can be found in the Appendix.

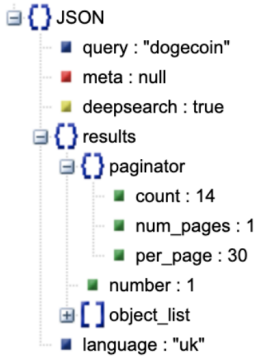


FIGURE 3.1: Declarations API simple JSON response

3.2 Data modeling

The primary purpose of this work is application of graph theory for corruption profiling which required us to transform the data to a universal format and also make it suitable for graph representation. First, we processed around 40 GB of text data from the complete archive of available electronic declarations from 2011 to 2020. Unquestionably, we sampled the data to shorten the sizes we work with. We decided to use only declarations from 2019 as this is the most recent year for which declarations are fully available by the time we are writing this paper. Also, to have a single declaration per person, we decided to use annual declarations - the ones that every professional submits at the end of each year. There are also different declaration types (e.g., before- and after dismissal or before-hire), but the argumentation for choosing annual declarations is simple: these declarations cover all the public servants, and they are well-standardized. In total, we found 1266593 declarations that matched our filters: they were annual NACP declarations from 2019 that were not 'corrected' by the declarant after submission (this way, we ensure we use final versions).

Next, we need to define a 'relationship' between two declarants to use this concept later for social graph modeling. To do this, the first thing we did was identifying the characteristics that could be markers of a possible connection between two public figures based on their declarations. When analyzing the declaration contents, we found two extractable entities that could mark a link: relation via a company co-ownership and relation via an organization. For example, if one public servant is a member of the same NGO with another public official, we assume that these two people are connected through the NGO.

For connections with companies, we used chapters 7, 8, 9, and 15 of a declaration: this means if a declarant has securities (7) or corporate rights (8) of a company,

they or their family members are controllers of a company (9), or they work for a company as their second job (15) - we claim that a relationship exists between the declarant and the company. For organizations, we used chapter 16 of the declaration: if a declarant is a member of an organization or its body, we claim the relationship between the declarant and the organization.

It is essential to mention that by no means all the declarants have relationships with companies or organizations. Out of the 1.2M annual declarations from 2019 that we found, only 9.4% had at least one connection with the organization or company (i.e., 119097 unique declarations). From there, 39% (46478) have connections only with a company/-ies, 54% (64216) have connections solely with an organization/-s, and 7% of declarations (8403) have links to both. Table 3.1 shows the shares of e-declarations in our final data set by the relation type they have.

Relation type	**Unique declarations**	**Declarations share of total**
7 Company - securities	14593	12.3%
8 Company - corporate rights	13552	11.4%
9 Company - beneficial owner	9482	8.0%
15 Company - concurrent job	24003	20.2%
16 Organization - member	72619	61.0%

TABLE 3.1: Number of unique declarations and their share in the final data set (annual 2019 NACP declarations) by type of relationship the declarant has with companies or organizations

We also did another piece of data cleaning and processing, such as excluding records of companies with empty or wrong EDRPOU [1] codes, filtering duplicate declarations that have different IDs, adding additional descriptive features, performing data transformations, and more. After that, the size of our final data set is 113 thousands unique declarations.

Our resulting data set primarily contains all the pairs 'company - declarant' and 'organization - declarant' that we were able to detect. The data attributes include information about the declarant and, depending on the relationship type, information about the company/organization that is a part of the pair. Figure 3.2 shows a snapshot of the final companies data set we built (some attributes are hidden for clearance). The two highlighted fields represent the declaration ID (i.e., person's ID) and the company ID. These fields are unique identifiers of people and companies, and they will be later used as our network nodes.

We provide full open source access to data collection, preparation and transformation steps. The complete Python script for data set creation from archived declarations data, along with the documentation for the data attributes, can be found in the project public repository on GitHub, under the `/tree/main/DataModelling` (ipopovych, 2021).

[1] EDRPOU - Ukrainian state registry legal entity identifier.

id	full_name	office	position	organization_group	relation_type	company_id	company_name
nacp_4a	Рогович Ол	Коломи	Начальник	Місцеві адміністрац	15	37446383	КЗ КРР "КРЦ ПМСД"
nacp_af1	Дубинська	Раківчиц	Депутат сі	Місцеві адміністрац	15	02228411	Відділ культури Кол
nacp_39	Павленко В	Ізмаїльс	Депутат Із	Місцеві адміністрац	15	26569293	Ізмаїльська міська р
nacp_56	Мельник Ю	Хмельни	Голова ра	Місцеві адміністрац	8	35344466	Приватне підриємс
nacp_d2	Іщенко Вал	ТОВ Дом	бухгалтер	Без категорії	15	03563548	ПРАТ ЧОП Агротехс
nacp_cc	Шевчук Євг	ТОВ "АВ	Менеджер	Без категорії	7	00381479	ВАТ "Дніпропетров
nacp_7b	Нещимний	КП ШКЗ	Помічник	Без категорії	7	00191158	ПрАТ МК Азов стал
nacp_85	Зубіцький С	Приват	Приватний	Без категорії	15	04403025	Держнянська місь
nacp_d5	Дейко Петр	ТОВ Теп	Головний	Без категорії	7	14085922	ПрАТ "Теплоенерге
nacp_d5	Дейко Петр	ТОВ Теп	Головний	Без категорії	8	42509827	ТОВ "ТЕПЛОЕНЕРГС
nacp_d5	Дейко Петр	ТОВ Теп	Головний	Без категорії	7	30965655	ВАТ "Сан Інбев Укр
nacp_51	Перезва Ол	Депутат	Депутат Н	Місцеві адміністрац	15	38225250	ПП "Чорноморець-
nacp_d5	Мєшкова Ін	Печерськ	головний	Кабмін, міністерства	15	30109129	ТОВ "Юрія-фарм"
nacp_8a	Дітяр Надія	Головне	головний	Інші державні служб	7	05747991	ВАТ СМНВО ім. М.В

FIGURE 3.2: A sample snapshot of the final data set of 'company - declarant' pairs. Some attributes (columns) are hidden in this picture

Chapter 4

Social networks of Ukrainian public officials and their characteristics

Empirical analysis of social networks and the relationships between the network members for corruption risk detection requires sufficient network visualization and descriptive analysis in the first place. This chapter presents the different approaches we developed to build network graphs of Ukrainian civil servants and discusses few examples of the obtained results.

The approach we suggest for linking nodes in the graph is linkage via company or linkage via organizations, which means that we have three types of nodes: declarant, company, organization, and two types of connections - declarant - company and declarant - organization (Figure 4.1).

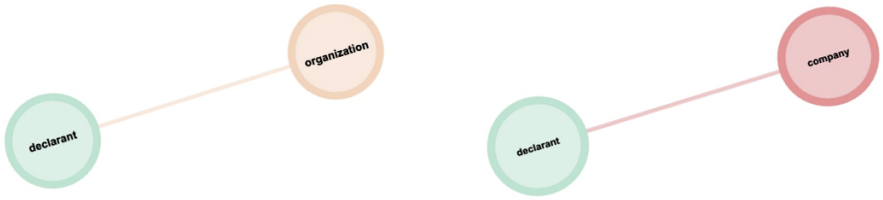


FIGURE 4.1: The three types of nodes and the two types of relationships we define in the network graph

An important thing to mention, due to the computation power limit, it was not possible to work with the entire data set at once efficiently. The final data set consists of 46302 unique companies, 17526 organizations, and 113052 people. If we were about to build a network graph based on a complete data set, the graph would consist of 176880 nodes. Such a graph is hardly visualizable and not advantageous for the research. That has driven our decision to use sampled data to explain our conceptual approach clearly: we used subsets of declarants based on the type of their governmental institution. In total, there are 19 types of institution groups (Appendix C).

4.1 Topological characteristics of network graphs

4.1.1 Declarant-company and declarant-organization graphs

We chose to focus on two groups of people: the Prosecutor's office and Ukrainian Parliament, because these two groups are similar in size and are not too large for meaningful visual graph representation. We modelled two types of networks based on 'declarant-company' and 'declarant-organization' relationships. The full network graphs visualizations can be found in Appendix D. Table 4.1 represents main characteristics of the resulting network graphs.

	declarant-company		**declarant-organization**	
	Parliament	Prosecutors	Parliament	Prosecutors
Number of nodes	2337	1437	1000	2752
Number of edges	1987	895	761	2653
Average degree	1,700	1,246	1,522	1,928
Network density	0,001	0,001	0,002	0,001
Modularity	0,983	0,995	0,968	0,882
Connected Components	351	548	255	185

TABLE 4.1: Main topological characteristics of the declarant-company and declarant-organization graphs

Analyzing the difference between company and organization networks for the two groups, we can see that the prosecutor declarant-organization network has the highest average degree. The reason for this is that this network has nodes with very high degrees: these are trade unions, organizations with lots of members. The nodes with high degrees are distinguishable on the prosecutor's organization-company network visualization (Appendix D.4).

All the networks have low network density, which means that the number of existing relationships in the graphs is very far from the possible number of relationships. The high modularity coefficient suggests that the networks are easily splittable into components (smaller communities). However, the size of the communities is more relevant for us than the number of them. We applied one of the existing state-of-the-art approaches for community detection in network graphs (Blondel et al., 2008) to check the sizes of connected components. Community size distributions (Figure 4.2) for all these four networks have visualizable outliers - the count of communities is large for communities of size 2. It shows that most of the networks' communities consist of 2 nodes which is just a single connection between a person and a company. Network visualizations also demonstrate that.

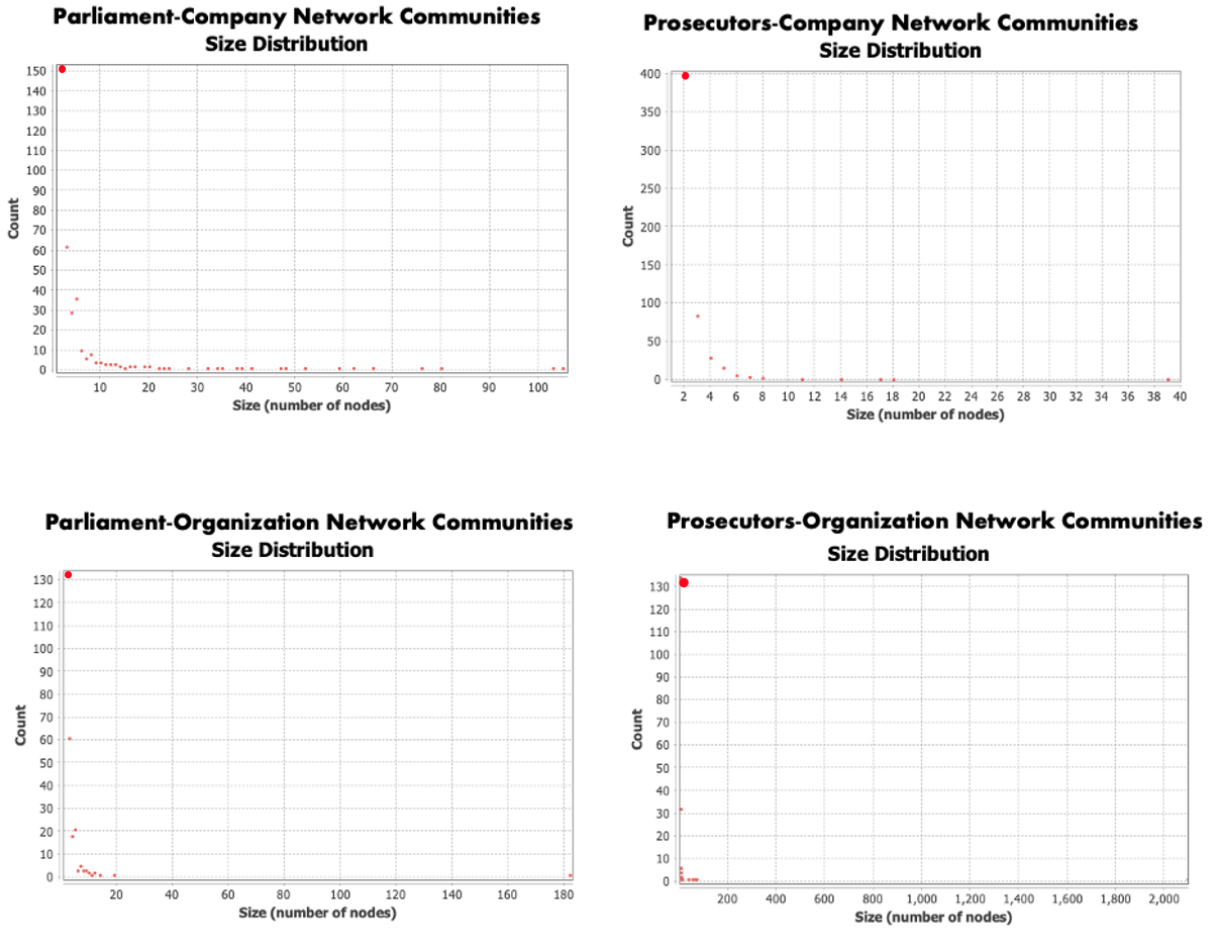


FIGURE 4.2: Community Sizes Distributions for 'declarant-company' and 'declarant-organization' graphs of Parliament and Prosecutors.

4.1.2 Declarant-company-organization graphs

Our next step was building graphs of people, companies, and organizations. The primary reason for combining the graphs is to track non-direct relationships among declarants: for example, two people can have a connection via a company, third person, and organization (Figure 4.3).

FIGURE 4.3: The indirect connection between declarants, via company and organization.

We performed a composition (i.e., a union of all nodes and edges) of the graphs described in section 4.1.1 to achieve the complete graph of all declarants who are members of parliament or the Prosecutors' office. This graph contains 4004 nodes, with an average degree of 1,77.

4.2 Reduction to declarant-declarant graphs

Our final goal is to analyze connections between people, so we decided to simplify the network graphs by removing all the nodes except the declarant nodes. We need to ensure we retain all the connections among people, so we can not merely drop all the irrelevant nodes and the adjacent edges out of the network, but we need to remove nodes and add a new edge instead if the removed node was a connecting node for two people.

Here is the approach that we used for this: first, create a list of all declarants nodes from the composed declarant-company-organization graph (Chapter 4.1.2); then, create a new network graph of people, add all these nodes and no edges so far; after that, iterate through all the node pairs and check if a path of length 2 exists between them in the declarant-company-organization graph. If a path exists, add an edge to the new network graph (Figure 4.4).

```
for node1 in people_nodes_graph.nodes():
    for node2 in people_nodes_graph.nodes():
        if ((node1 != node2) & (nx.has_path(parent_graph, node1, node2))):
            shortest_path_length = nx.shortest_path_length(parent_graph, node1, node2)
            if shortest_path_length <= 2:
                final_graph.add_edge(node1, node2)
                final_graph.edges[node1, node2]['path_length']= shortest_path_length
```

FIGURE 4.4: A snippet of Python code for graph simplification to a person-person graph. `people_nodes_graph` is a list of all person nodes, `parent_graph` is a composed declarant-company-organization graph.

For shortest path length calculation, we used NetworkX (Hagberg, Schult, and Swart, 2008) implementation of Dijkstra's (Dijkstra, 1959) shortest path algorithm is used. We use the threshold for path of length 2 to only cover the direct relationships, like person-company-person or person-organization-person and do not include indirect ones (Chapter 4.1.2), for which the path length will be 4 or more. The approach is flexible, so this could be adjusted to whatever threshold for the maximal path length between two nodes to consider they are connected.

We have also added various attributes of the graph nodes for latter use in the analysis and visualization: their office, position type, and location. Figure 4.5 shows the visualization of the final person-person graph of Parliament members and the employees of Prosecutors' office, colored by their location.

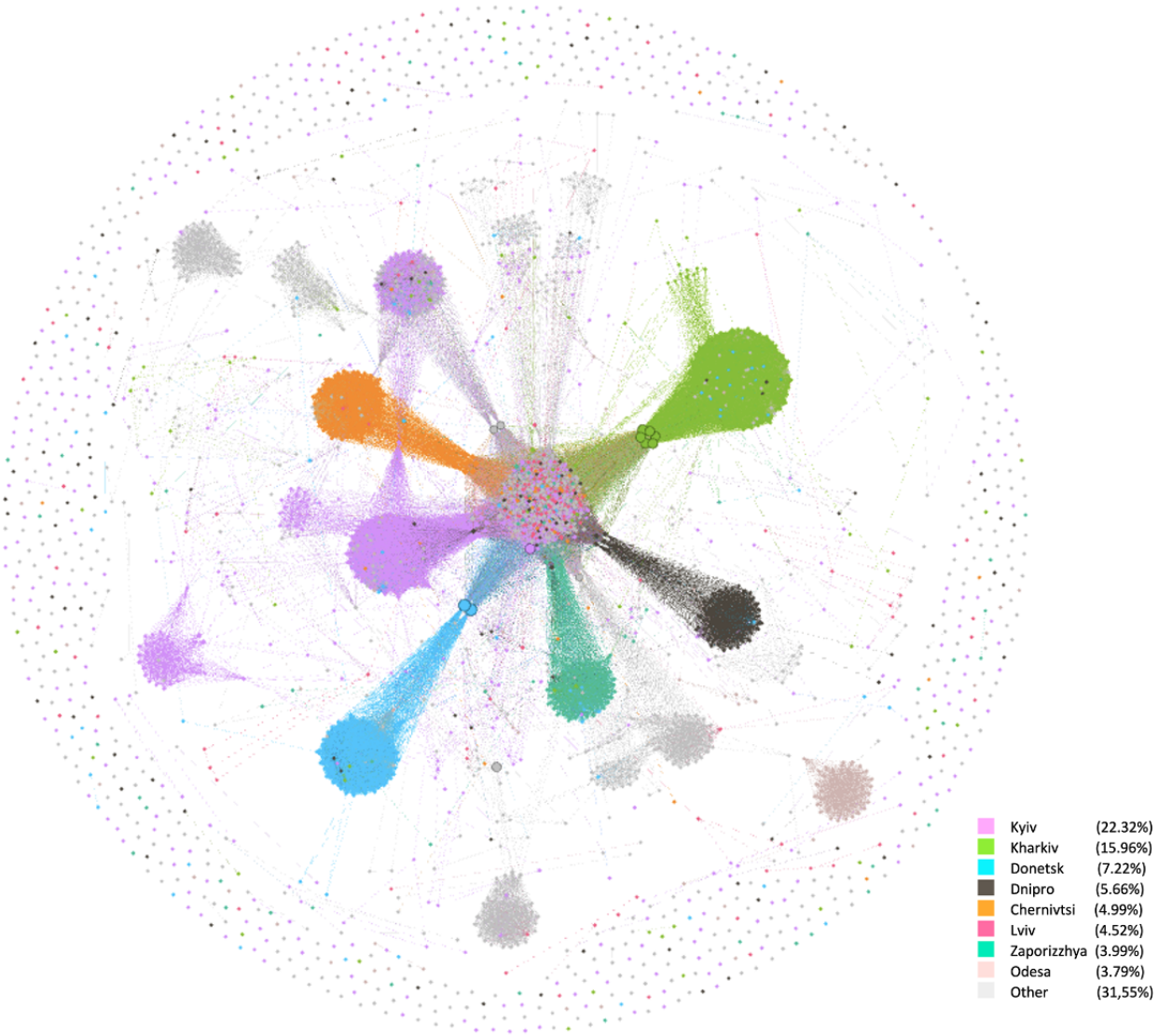


FIGURE 4.5: The person-to-person relationship graph, colored by location of the declarant

Chapter 5

Application of networks for corruption risk profiling

5.1 Important features of declarant-declarant graphs

The peer-to-peer graphs that we built are a powerful tool for investigating social connections. This section explains some key features of these graphs. One insightful finding from the declarant-to-declarant network graph (Figure 4.5) is that declarants were 'naturally' grouped into connected components of people who work in the same cities by our network construction algorithm, even though we did not use the office location as a parameter for modeling. This example network shows us that there are strong communities of public servants in some cities (e.g., Kyiv or Kharkiv), while for other cities, like Lviv, we do not observe distinct connected subgraphs.

Another significant metric is betweenness centrality of the nodes. Nodes with high betweenness centrality are the nodes that connect different communities. Burt (Burt, 1992) has studied that the network members who are bridging "structural holes" (gaps between communities) are powerful, because they exercise control (e.g., decide whether to share information or not) over the persons they connect between. Figure 5.1 shows a snapshot of the prosecutors-parliament network, where the size of the node corresponds to its betweenness centrality. Some network 'connectors' can be observed visually because of their large node sizes.

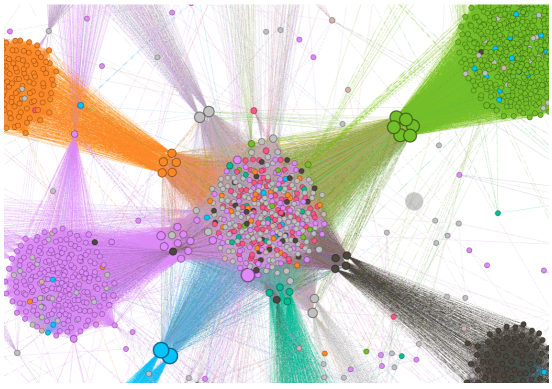


FIGURE 5.1: A snapshot of the nodes with high betweenness centrality. The size of the node indicates its betweenness centrality

One more crucial feature that we can track from this graph is the connectivity between different offices: we can spot the connectors between Kyiv and Kharkiv prosecutor's offices or find people who link parliament communities to the prosecutor subgraphs.

Figure 5.2 represents a person-to-person graph of people from Court, Governmental commissions, Prosecutor's office, Pension Fund and Parliament. When building this graph, we excluded all companies and organizations that had node degree larger than 30 from the parent graph. We did this to ensure we are not linking people who are members of some big organizations and do not actually have a link in real life. As we can see, even after we excluded large connective companies and organizations, the person-to-person links still remain in a graph, smaller communities are detectable, and we can observe the links between different communities.

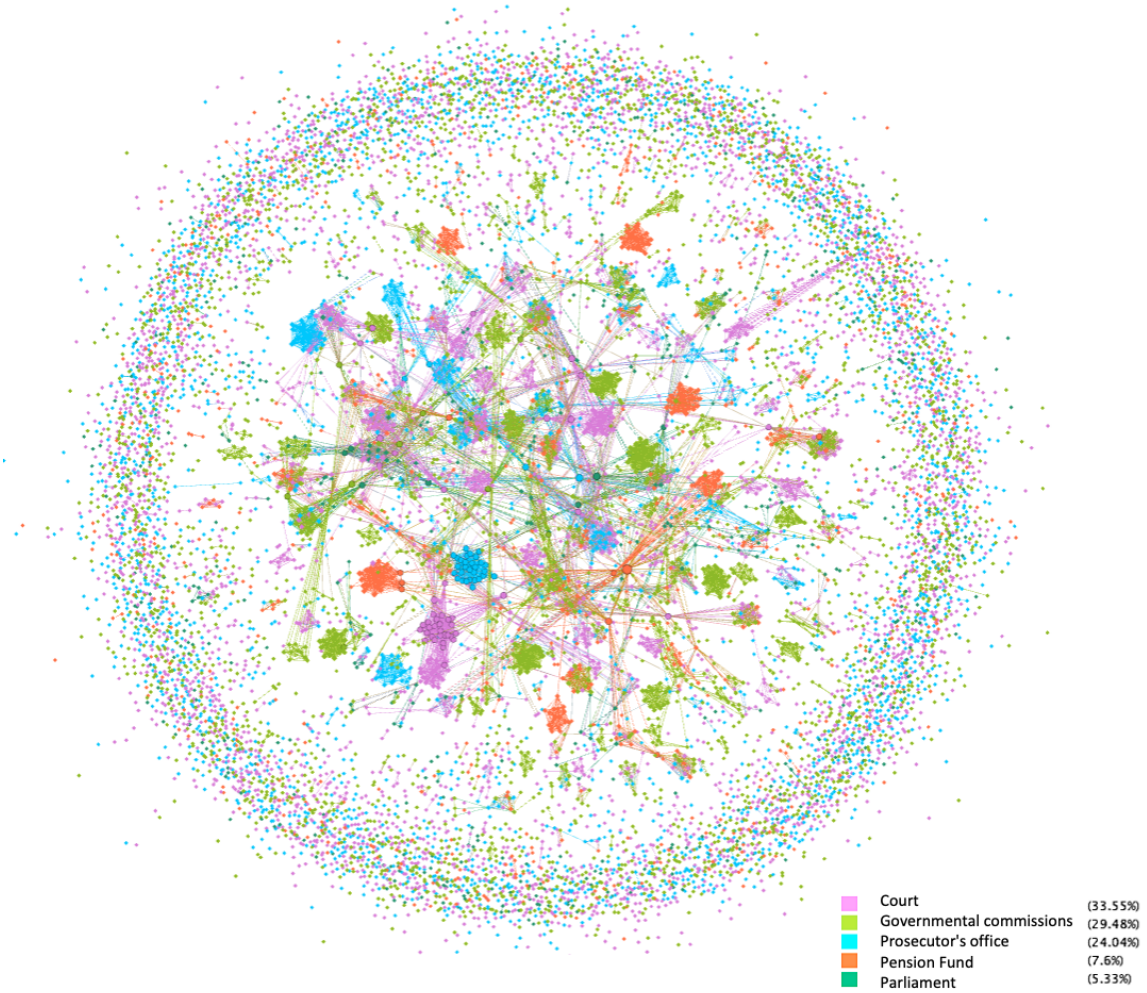


FIGURE 5.2: A declarant-to-declarant network graph of people from four different office types.

5.2 Possible ways of corruption risk profiling using network graphs

Our final step in this research is a proposal of an approach for corruption risk profiling. After analyzing and modeling the USRD data, we can conclude that the data we have in the declarations open dataset is not enough to predict if a governmental worker committed corruption. However, we can eliminate some strong indicators of a person's level of influence in the social network graph, which can be a good proxy for corruption risk. Also, there is an opportunity to extract other features that could potentially be useful for corruption detection from the declarations' dataset itself. For example, we can track a person's assets over time and catch all the significant increases or decreases in the total amount of assets. Some more declaration risk factors can be designed and defined: the total wealth of the declarant's family, luxury cars, and others. One more opportunity is to build weighted graphs based on numbers of different connections between people and track strong relationships this way. Our suggestion is either to work on the graph analysis results with some corruption domain experts or to use these network indicators in combination with data from other sources.

One of the suggestions for external data is The Unified State Register of Persons Who Have Committed Corruption or Corruption-Related Offenses [1], which was launched recently in 2019. It is a portal that contains information on all individuals and legal entities that have committed corruption offenses. It does not have sufficient data yet, but we highly recommend it for future research. The register contains information about individuals who have committed corruption or corruption-related offenses and also legal entities to which criminal law measures have been applied (for example, a fine or confiscation of property) due to a corruption offense. We can use this information to analyze corrupt officials to identify areas of public policy and positions with the highest corruption risks. This means that we will be able to mark all the 'corrupted' nodes in our public official's social network and potentially track the corrupted sub-graph of people connected to them by using it.

[1] Accessible via NACP website.

Chapter 6

Conclusions

6.1 Research summary

In this work, we focused on three main goals: use Ukrainian open data for empirical research and analysis, model networks of Ukrainian public servants, and investigate the possible approaches for scoring the corruption risk of a network member. To achieve the goals, we first analyzed the possible data sources and chose declarations.com.ua API as the best option, primarily because of its stability and access to larger information volumes. After analyzing and processing the USRD data, we suggested two possible approaches for network modeling by defining two types of connections between public officials: connection via company co-ownership and connection via an organization. We built some primary networks based on subsets of data and discovered that these networks have a low average degree and low density. However, the modularity coefficient for most of them is close to 1. We discovered that this primarily happened due to the networks containing lots of disconnected 'company person' and 'organization-person' components. That has driven our decision to perform network simplification and build person-to-person networks. Analysis of the structures of these networks showed us that they help detect communities and people with high betweenness centrality. We observed communities that link people from the exact locations (Figure 4.5), as well as office-based communities (Figure 5.2) and communities of political party members.

Finally, we suggested a set of methods that can be used for future corruption risk profiling with graphs. We defined three main directions for corruption risk detection analysis: adding more insightful node features to the network graphs from the current data set, working with corruption domain experts to define corruption-specific network characteristics, and using external data to expand the networks data set with more indicators, for example, corruption committed in the past. We think that a combination of these three approaches could help to reach better results.

6.2 Future work

A wide range of possible future research derives from our work. One of the directions is a deeper study of graph types and graph features. This includes modeling dynamic graphs based on the time series characteristics of the networks and their changes. Another possible research idea is a study of graph statistics for subgraphs of public servant networks groups, their analysis, and the differences among groups. Besides that, a conceptually interesting idea is adding more types of connection edges to graphs. Some of these may include more formal relationships like common property, as well as informal relationships such as, for example, common university/army or co-authorship in science papers. Finally, graph-based anomaly detection can be applied to contribute to further research around corruption risk prediction for a node.

The main thing we learned is that the problem we addressed is a complex one, but it is an incredibly beautiful one at the same time. For years, these kinds of problems were approached separately by social scientists and by formal scientists, each looking for answers to their own questions. We believe that extensive collaboration and a multidisciplinary approach will help us find the truth somewhere in the middle.

Appendix A

Declarations API response structure.

For each declaration, API returns a document consisting of five sections:

- infocard — a short card of the document describing the declarant (full name, place of work, and position), type of declaration, etc.
- guid (id) — the document's unique string code
- raw_source — for old, paper declarations, it will have a full machine-readable declaration document in the old format. For new NACP declarations, it will include a link to a machine-readable copy of the declaration on the NACP's website
- unified_source — for old declarations, it will have a declaration that has been converted from the old format to the new one, similar to the NACP's. For new NACP declarations, it will include a complete new declaration document with small format adjustments.
- related_entities — contains information on the related natural and legal persons of the declarant, as well as documents. Natural persons include the declarant's family. Legal persons include those companies in which the declarant is a beneficiary (owned by, only the EDRPOU code), has a financial connection (related, only the EDRPOU code), as well as all the companies mentioned in the declaration (all, may yield both EDRPOU codes and names). Finally, linked documents contain references to existing corrected (or original) declarations of the current one.

Source: declarations.com.ua

Appendix B

Overview of the declaration structure.
Every standardized declaration contains 16 chapters:

1. Type of declaration and the reporting period
2. Information about the declarant
    - Information about the declarant
    - Information about the declarant's family members
3. Immovable property
4. Objects of unfinished construction
5. Valuable movable property (other than vehicles)
6. Valuable movable property - vehicles
7. Securities
8. Corporate rights
9. Legal entities, trusts, or other similar legal entities, the ultimate beneficial owner (controller) of which is the declaring entity or his family members
10. Intangible assets
11. Revenues, including gifts
12. Cash assets
    - Banking and other financial institutions holding accounts of the declaring entity or his family
13. Financial obligations
14. Expenses and transactions of the declarant
15. The concurrent job of the declarant
16. Membership of the declarant in organizations and their bodies

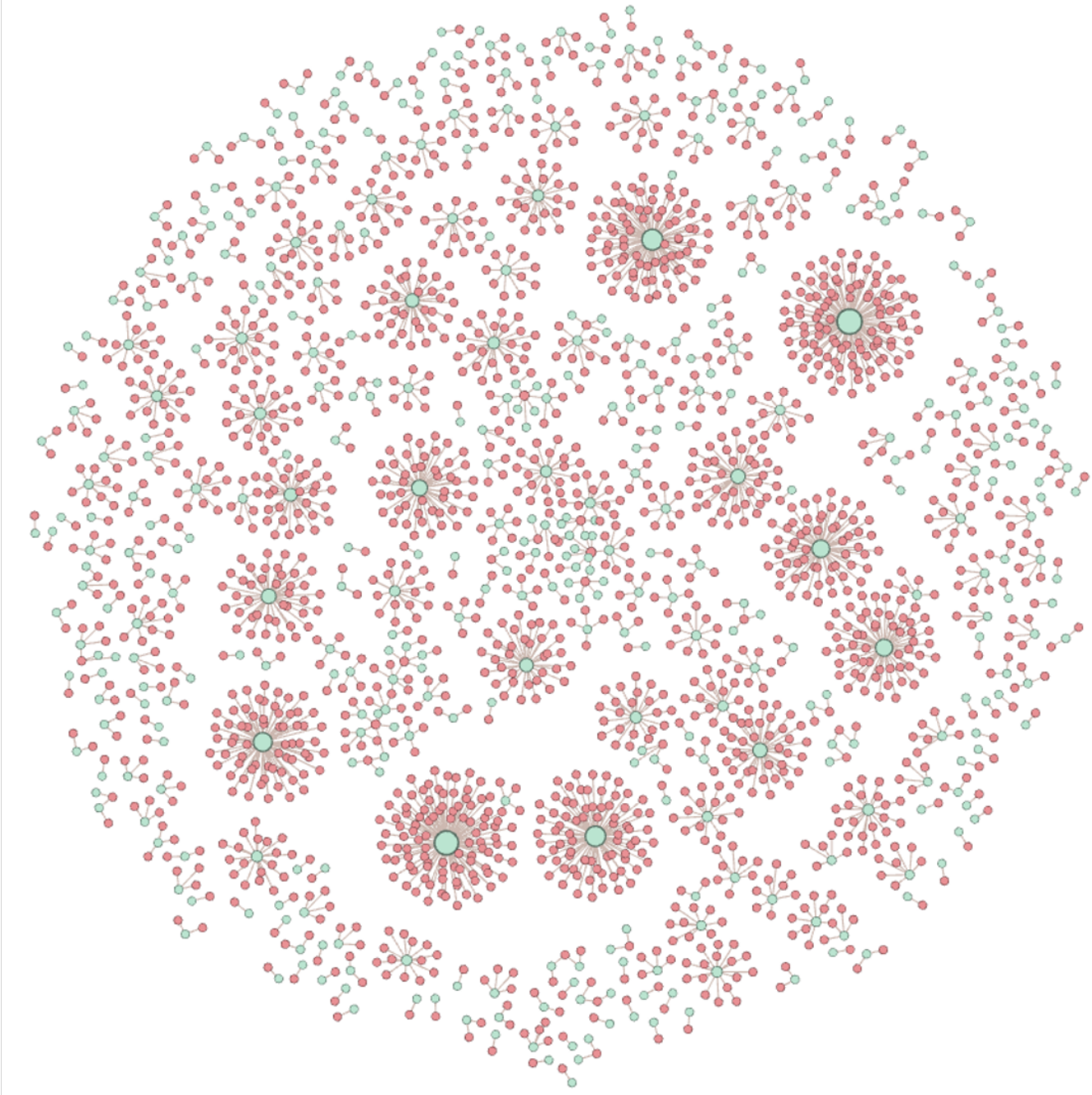
Appendix C

The groups of governmental positions and number of people by the type of their position

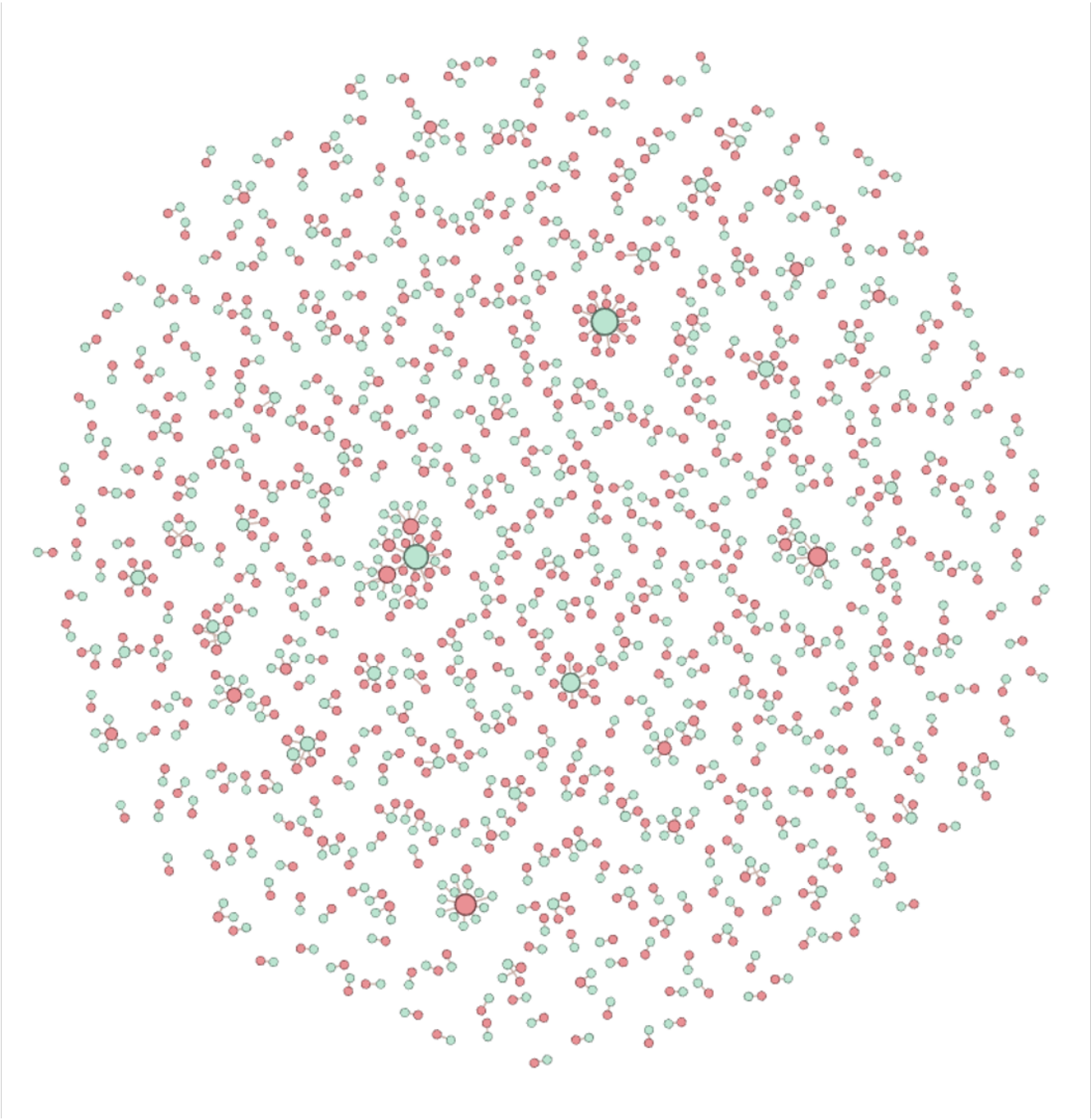
Organization Group	# of people
Місцеві адміністрації та ради	17814
Без категорії	15729
Кабмін, міністерства та підлеглі органи	8482
Суд	2145
Інші державні служби, комісії, і т.п.	1779
Прокуратура	677
Пенсійний фонд	673
Парламент	426
НБУ	240
Фонд державного майна	139
Антимонопольний комітет	80
Адміністрація / Секретаріат Президента	65
Рахункова палата	63
НАБУ	52
НАЗК	34
Державний комітет телебачення і радіомовлення	27
Вища рада юстиції	27
ЦВК	24
СБУ	20

Appendix D

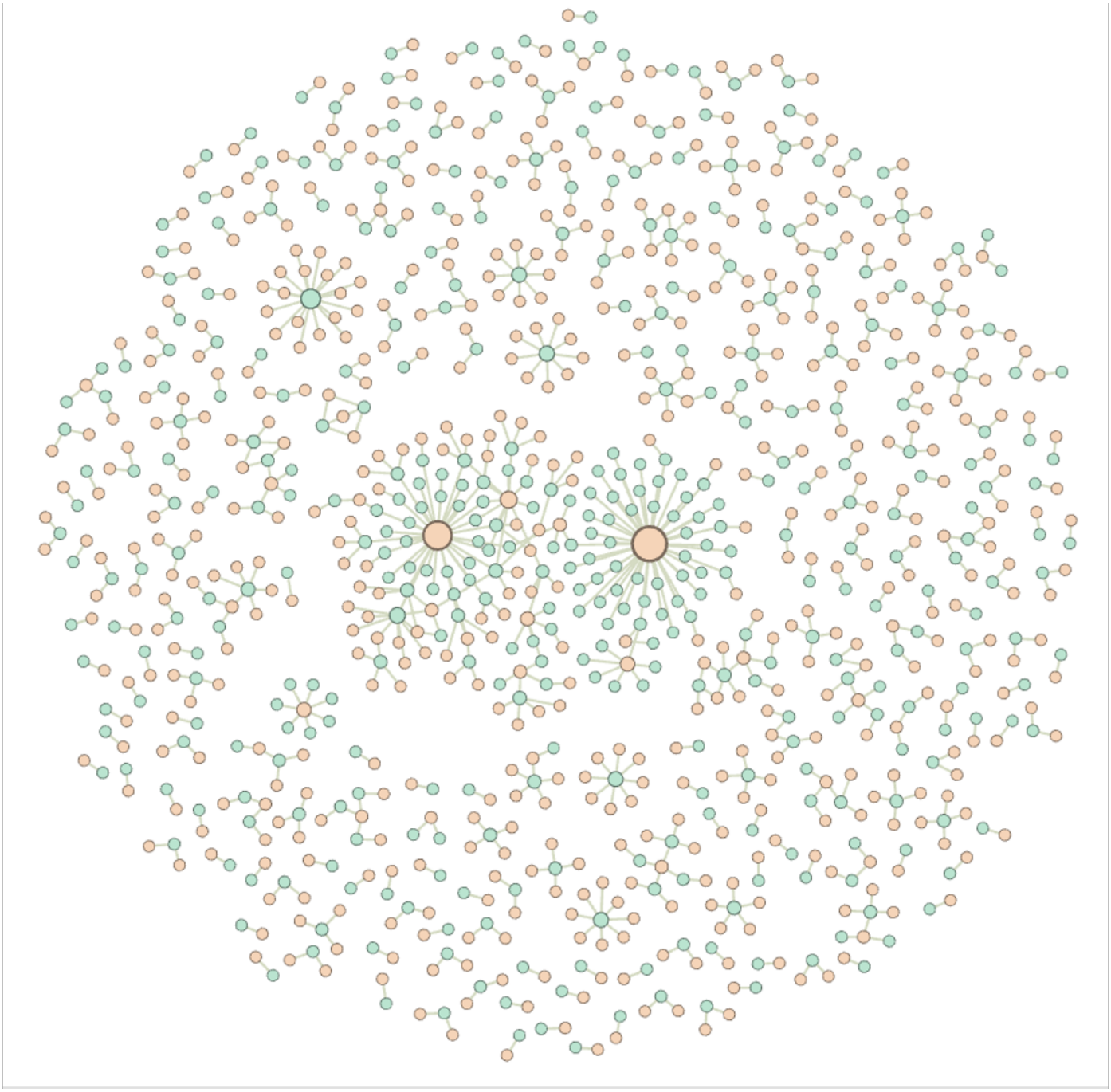
Parliament declarant-company network visualization.



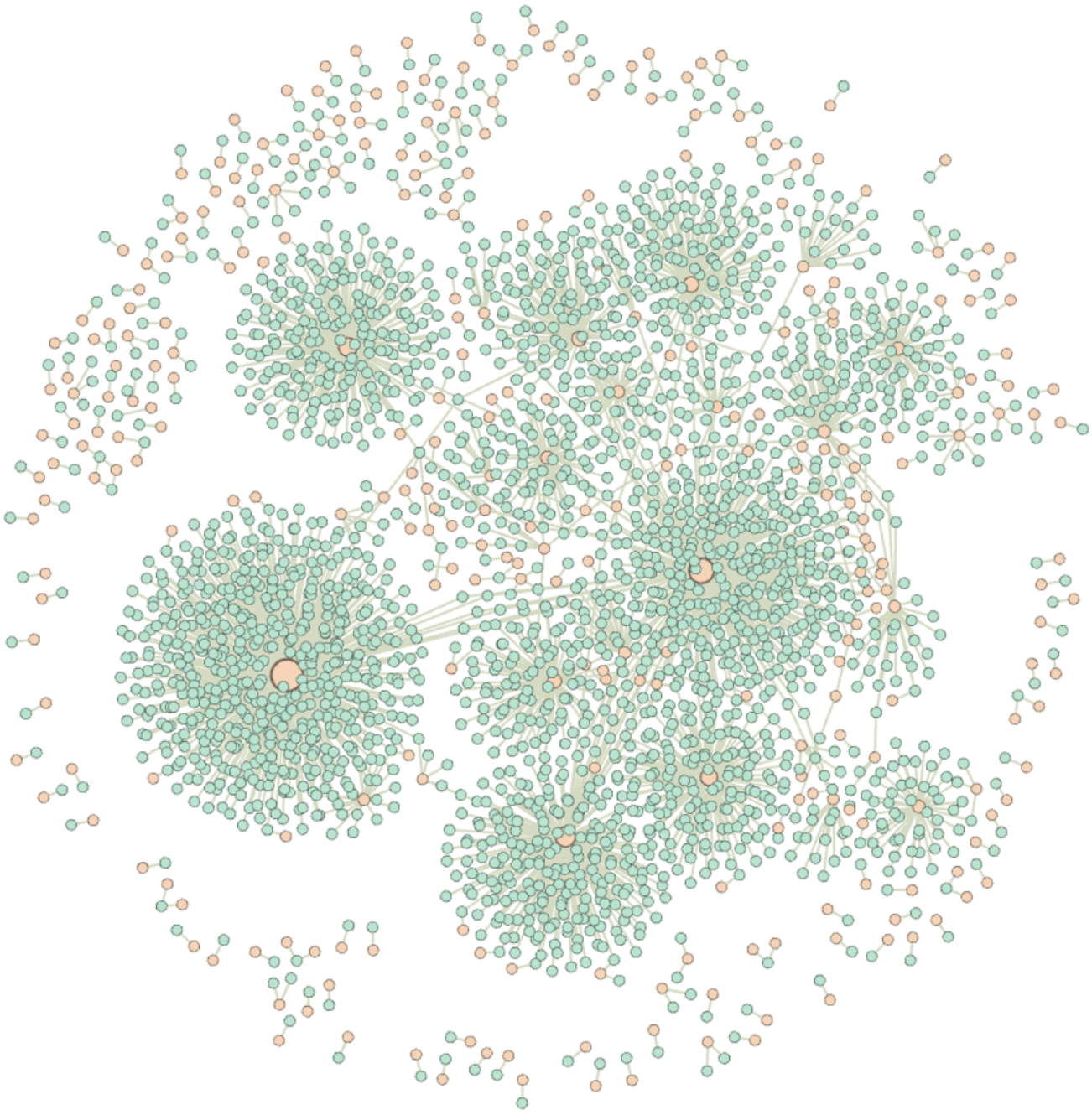
Prosecutors declarant-company network visualization.



Parliament organization-company network visualization.



Prosecutors organization-company network visualization.



Bibliography

Backstrom, Lars and Jon Kleinberg (Oct. 2013). "Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook". In: *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. DOI: 10.1145/2531602.2531642.

Blondel, Vincent D et al. (2008). "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/p10008. URL: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

Burt, Ronald S. (1992). *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

Can, Umit and Bilal Alatas (2019). "A new direction in social network analysis: Online social network analysis problems and applications". In: *Physica A: Statistical Mechanics and its Applications* 535, p. 122372. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2019.122372. URL: https://www.sciencedirect.com/science/article/pii/S0378437119313597.

Declarations.com.ua, website (2021). *Declarations.com.ua - the largest database of declarations of public officials in Ukraine, website*. URL: https://declarations.com.ua/en/ (visited on 05/16/2021).

Dijkstra, Edsger W (1959). "A note on two problems in connexion with graphs". In: *Numerische mathematik* 1.1, pp. 269–271.

Duguay, Raphael, Thomas Rauter, and Delphine Samuels (Nov. 2020). "The Impact of Open Data on Public Procurement". In: DOI: 10.2139/ssrn.3483868. URL: https://ssrn.com/abstract=3483868.

Hagberg, Aric A., Daniel A. Schult, and Pieter J. Swart (2008). "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, pp. 11 –15.

Herrera Murillo, D. J. (Sept. 2019). "Using social network analysis in open contracting data to detect corruption and collusion risks". In: URL: https://research.tue.nl/en/studentTheses/using-social-network-analysis-in-open-contracting-data-to-detect-.

ipopovych, GitHub (2021). *Open Data Networks - our project public repository on GitHub*. URL: https://github.com/ipopovych/open_data_networks (visited on 05/16/2021).

Johnston, Michael D. (1999). "Public Officials, Private Interests, and Sustainable Democracy: When Politics and Corruption Meet". In:

Kostiuchenko, Tetiana (2012). "Elite Continuity in Ukraine: When Networks Matter (?)" In: *Historical Social Research / Historische Sozialforschung* 37.2 (140), pp. 14–25. ISSN: 01726404. URL: http://www.jstor.org/stable/41636574.

Kudelia, Serhiy (Sept. 2016). "Corruption in Ukraine: Perpetuum Mobile or the Endplay of Post-Soviet Elites?" In: ISBN: 9780804798457. DOI: 10.11126/stanford/9780804798457.003.0004.

Mauro, Paolo (Aug. 1995). "Corruption and Growth*". In: *The Quarterly Journal of Economics* 110.3, pp. 681–712. ISSN: 0033-5533. DOI: 10.2307/2946696. eprint: https://academic.oup.com/qje/article-pdf/110/3/681/5261114/110-3-681.pdf. URL: https://doi.org/10.2307/2946696.

NACP, website (2021). *The National Agency on Corruption Prevention of Ukraine, Official Website*. URL: https://nazk.gov.ua/en/about-nacp-2/ (visited on 05/16/2021).

Open Data Handbook, website (2021). *What is Open Data?* URL: https://opendatahandbook.org/guide/en/what-is-open-data/ (visited on 05/16/2021).

OpenGovData, website (2007). *The Annotated 8 Principles of Open Government Data*. URL: https://opengovdata.org/ (visited on 05/16/2021).

Prozorro, website (2021). *Prozorro official website*. URL: https://prozorro.gov.ua/ (visited on 05/16/2021).

Punch, Maurice (Sept. 2000). "Police Corruption and Its Prevention". In: *European Journal on Criminal Policy and Research* 8, pp. 301–324. DOI: 10.1023/A:1008777013115.

Quah, Jon S. T. (2001). "Combating Corruption in Singapore: What Can Be Learned?" In: *Journal of Contingencies and Crisis Management* 9.1, pp. 29–35. DOI: https://doi.org/10.1111/1468-5973.00151. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-5973.00151. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-5973.00151.

Rumyantseva, Nataliya L. (2005). "Taxonomy of Corruption in Higher Education". In: *Peabody Journal of Education* 80.1, pp. 81–92. DOI: 10.1207/S15327930pje8001_5. eprint: https://doi.org/10.1207/S15327930pje8001_5. URL: https://doi.org/10.1207/S15327930pje8001_5.

Safarov, I., A. Meijer, and S. Grimmelikhuijsen (2017). "Utilization of open government data: A systematic literature review of types, conditions, effects and users". In: *Inf. Polity* 22, pp. 1–24.

Tanzi, Vito (1998). "Corruption Around the World: Causes, Consequences, Scope, and Cures". In: *IMF Staff Papers* 45.4, pp. 559–594. URL: https://ideas.repec.org/a/pal/imfstp/v45y1998i4p559-594.html.

The Open Data Institute, website (2018). *How Ukraine became an open data pioneer*. URL: https://theodi.org/article/how-ukraine-became-an-open-data-pioneer/ (visited on 05/16/2021).

Transparency International, website (2021). *Transparency International CORRUPTION PERCEPTIONS INDEX 2020*. URL: https://www.transparency.org/en/cpi/2020/index/ukr (visited on 05/16/2021).

Ubaldi, Barbara (2013). "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives". In:

Vian, Taryn (Mar. 2008). "Review of corruption in the health sector: theory, methods and interventions". In: *Health Policy and Planning* 23.2, pp. 83–94. ISSN: 0268-1080. DOI: 10.1093/heapol/czm048. eprint: https://academic.oup.com/heapol/article-pdf/23/2/83/1591087/czm048.pdf. URL: https://doi.org/10.1093/heapol/czm048.

Wederman, Andrew (2004). "The Intensification of Corruption in China". In: *The China Quarterly* 180, 895–921. DOI: 10.1017/S0305741004000670.