APPLIED SCIENCES FACULTY.BA

Corruption Risk Profiling With Open Data: Application Of Graph Theory

JUNE 2021

Author:
Iryna Popovych

Supervisor:
Liubomyr Bregman

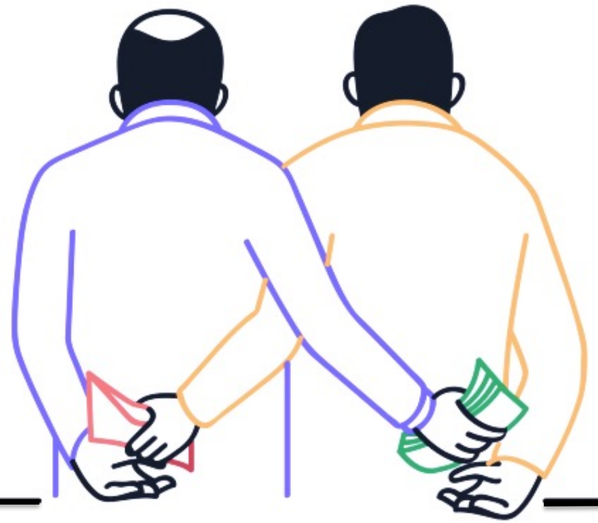
Agenda

- Introduction
- Data overview: USRD (Unified State Register of Declarations)
- Data modelling
- Social networks of Ukrainian public officials
  - Company and organization-based network graphs (two-mode networks)
  - Reduction to person-to-person network graphs (one-mode network)
- Analysis and Discussion on application of networks for corruption risk profiling
- Summary
- Q&A

JUNE 2021

Motivation 1: Corruption

- Corruption is a huge problem
- It is widely discussed by researches
- Political elites' networks – high corruption risk



Motivation 2: Open data

No decent academic discussion in the traditional research exchange

Big push for opening data, little progress in data monitoring and usage

There is often no final value for the society, which should be brought by usage

Our research setting

Lack of understanding of the circumstances under which corruption evolves.

Most of the research around corruption is using **descriptive approaches rather than quantitative methods**.

There are precedents that corruption develops inside communities of **political elites in Ukraine**.

No quantitative research that uses Ukrainian open governmental data to study social networks.

Our research setting and goals

Lack of understanding of the circumstances under which corruption evolves.

Most of the research around corruption is using **descriptive approaches rather than quantitative methods**.

There are precedents that corruption develops inside communities of **political elites in Ukraine**.

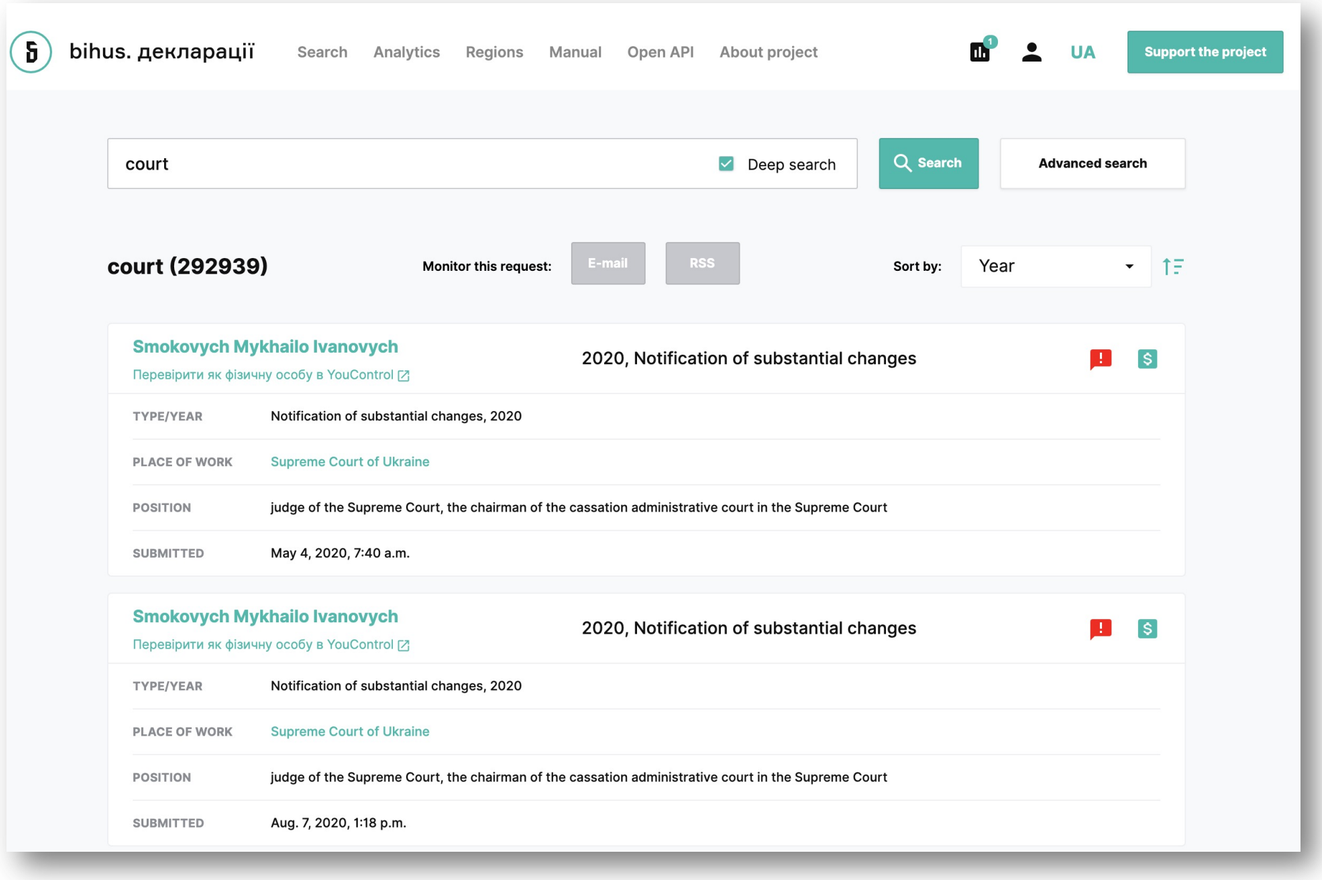
No quantitative research that uses Ukrainian open governmental data to study social networks.

Based on the information gained during subject exploration, we define our research goals as follows:

1. To study the feasibility and complexity of processing Ukrainian large scale open data and use it for empirical research and analysis
2. To analyse links between Ukrainian public servants and model social networks based on open data
3. To identify connected groups within networks and investigate the possible approaches for scoring the corruption risk of a network member

USRD data and declarations API

The motivation for choosing USRD data for this work is that we have seen researches on public procurement data and corruption, but **there is little or no research around USRD data.** Also, this is the **primary data set with information on political elites.**



Declarations.com.ua **benefits**:

- The largest database
- Provides stable open API access
- Well-documented
- Standardized JSON data

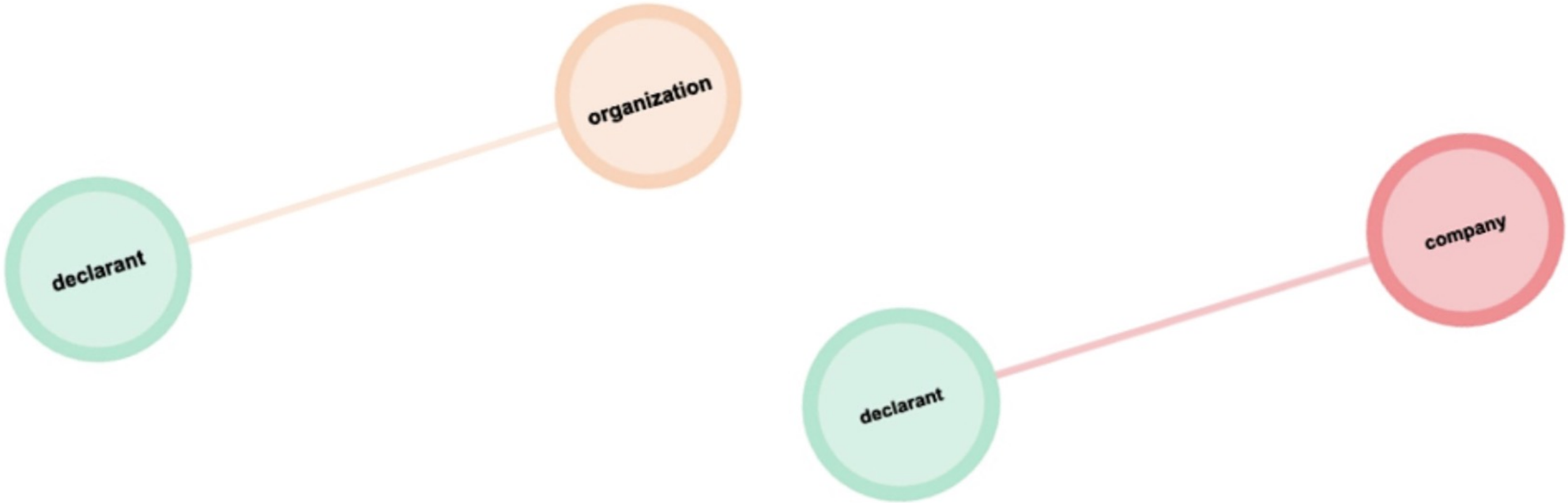
Data selection

1	2	3	4	5
Full declarations archive	2019 declarations	Annual declarations	Not corrected	With company/ organization
5 449 382 100%	1 299 617 24%	933 714 72%	854 986 92%	119 097 14%

Transformed data set snapshot

id	ull_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	Приватн	Приватний	Без категорії	15	04403025	Деражнянська міс
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 Не	Місцеві адміністрац	15	38225250	ПП "Чорноморець-
nacp_d5	Мєшкова Ін	Печерсь	головний	Кабмін, міністерства	15	30109129	ТОВ "Юрія-фарм"
nacp_8a	Дігтяр Надія	Головне	головний	Інші державні служб	7	05747991	ВАТ СМНВО ім. М.В

We define two relationship types

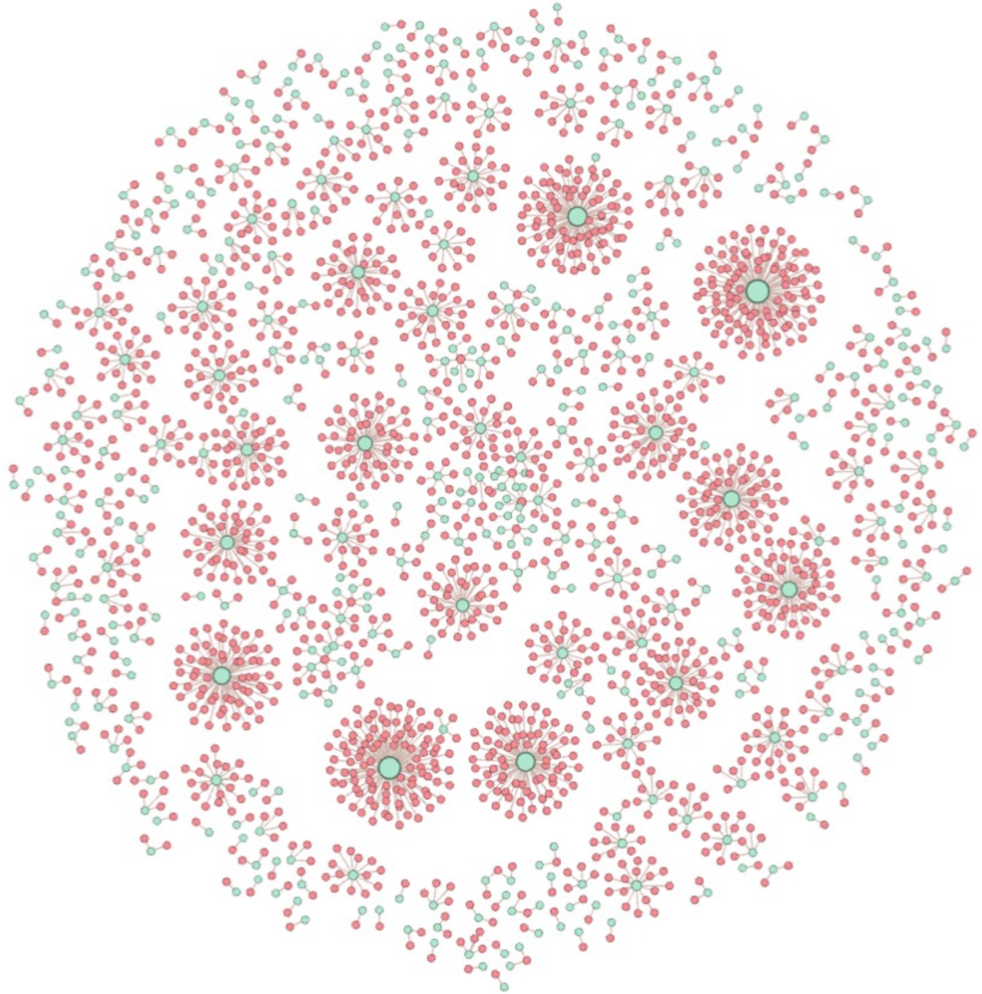


Declarations by relation type

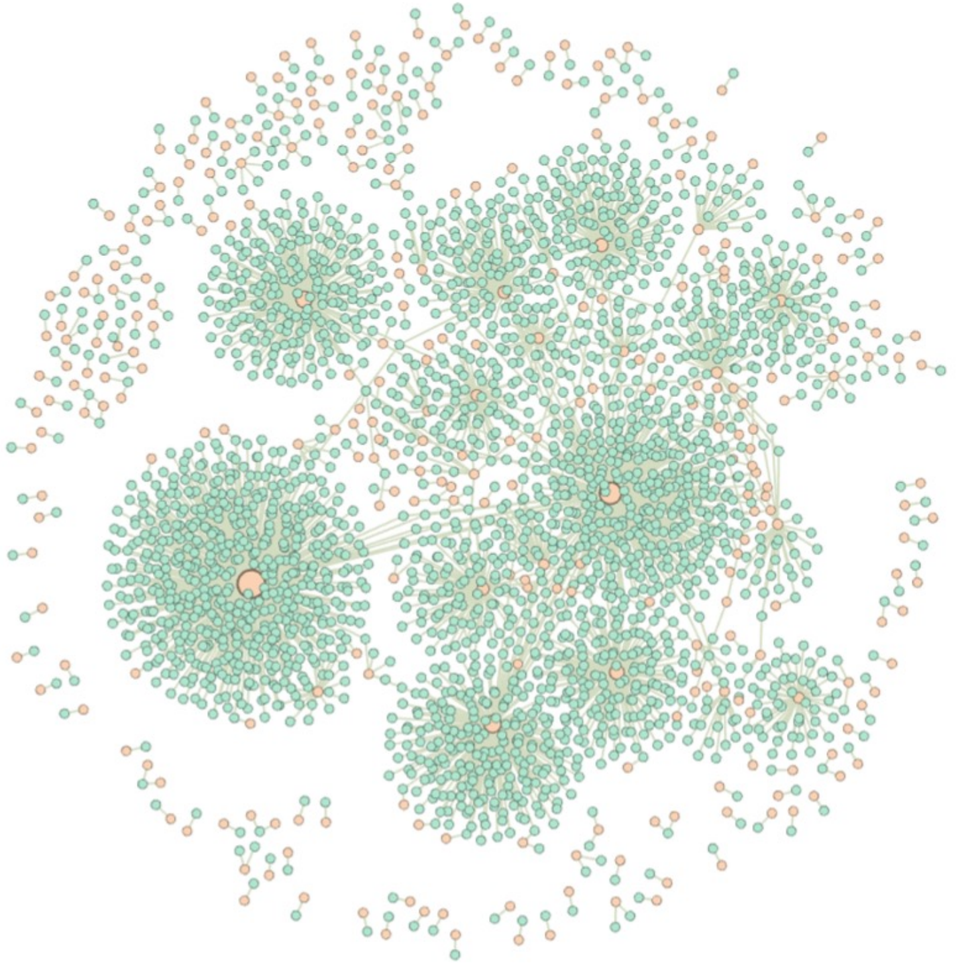
From the filtered declarations, 39% (46478) have connections only with a company(ies), 54% (64216) have connections solely with an organization(s), and 7% of declarations (8403) have links to both.

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%

Two-mode networks



Parliament declarant-company network visualization.

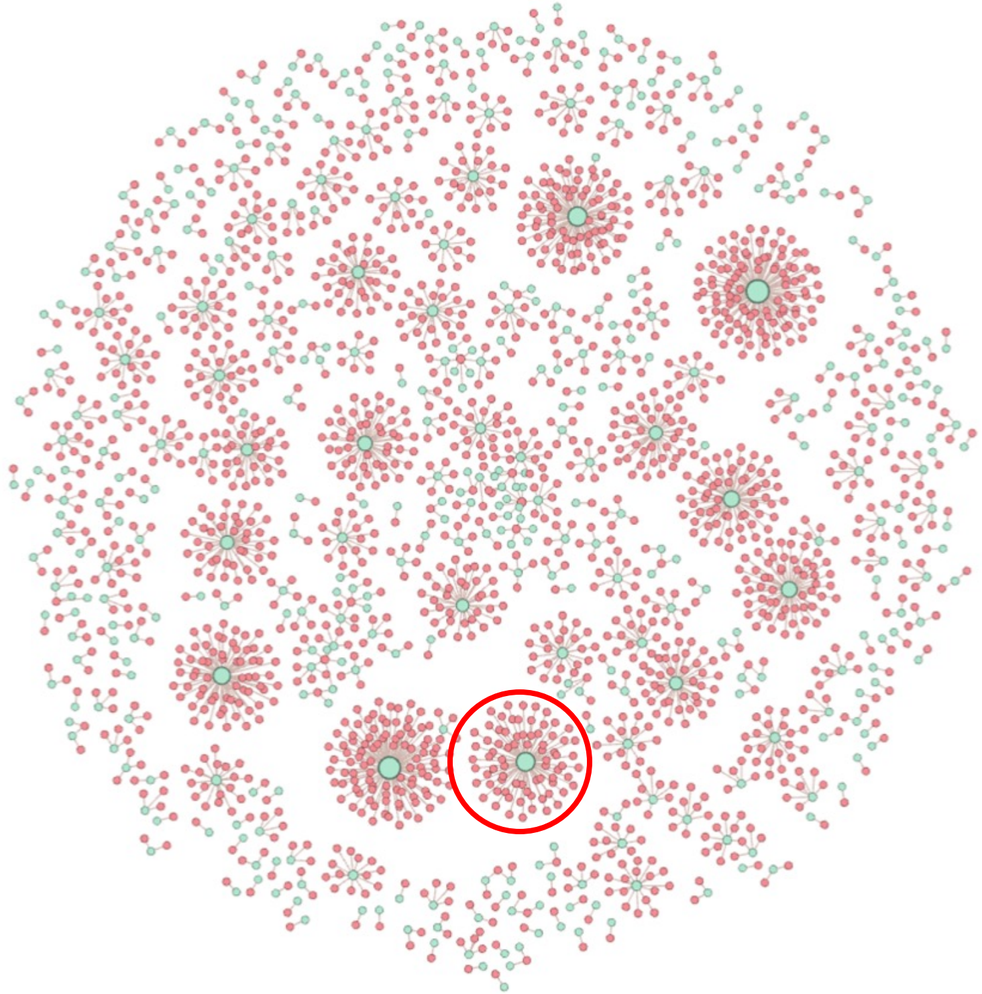


Prosecutors' declarant-organization network visualization.

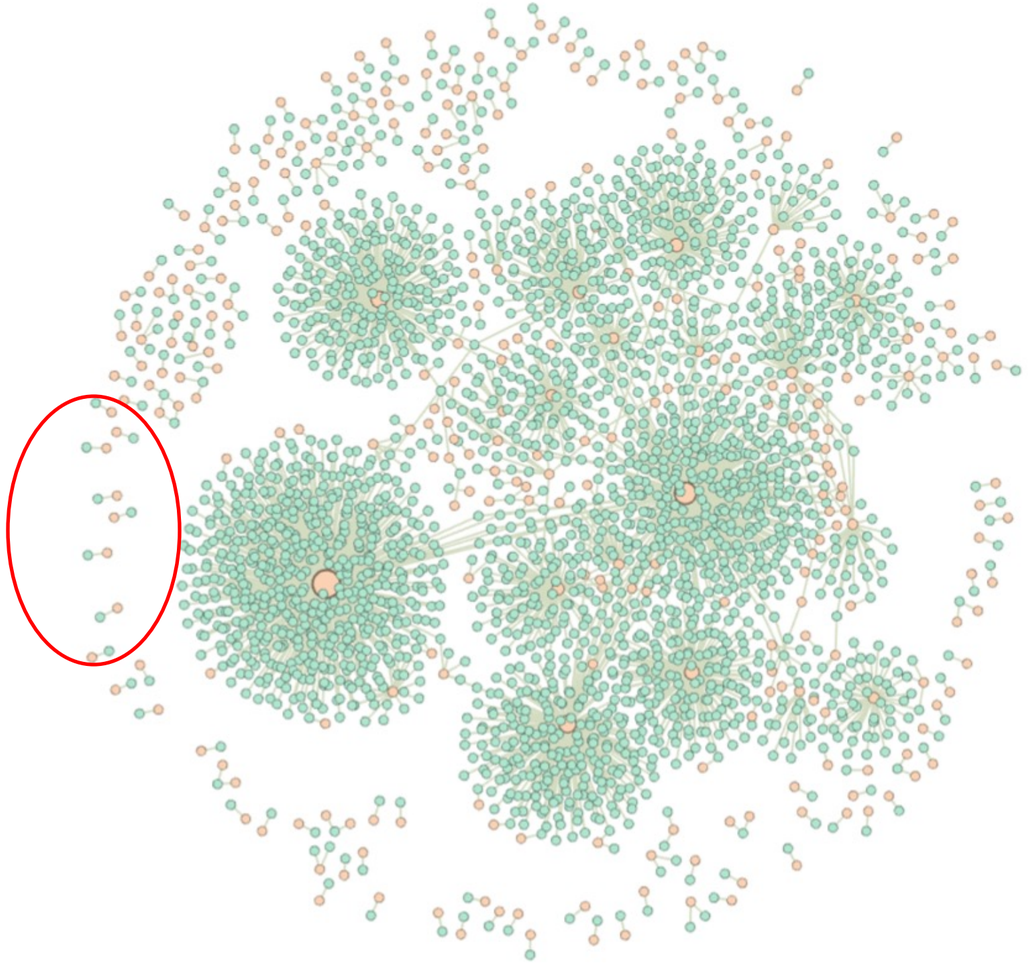
Two-mode networks characteristics: low average degrees and low density

	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

Two-mode networks issue: disconnected pairs

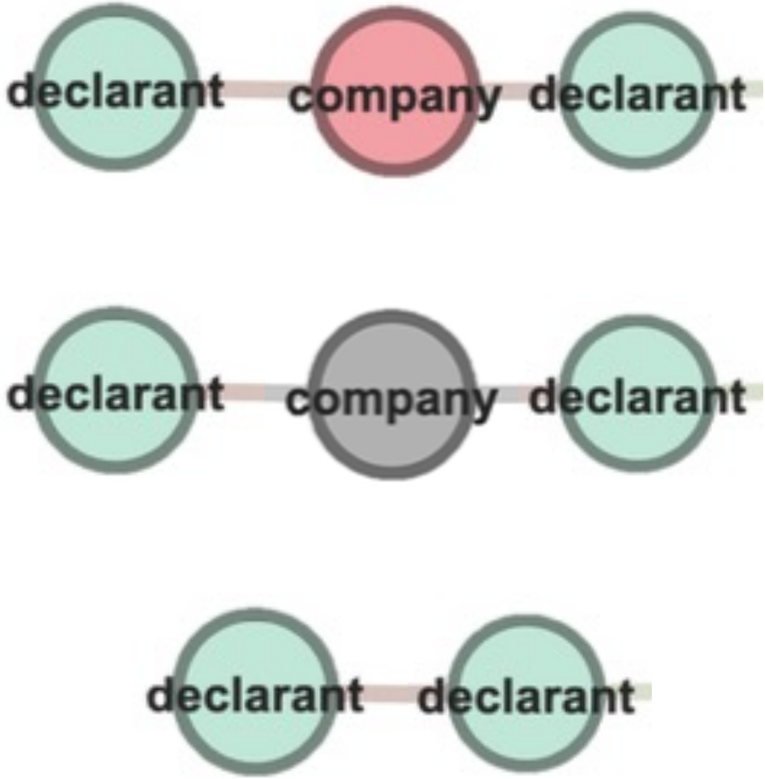


Parliament declarant-company network visualization.



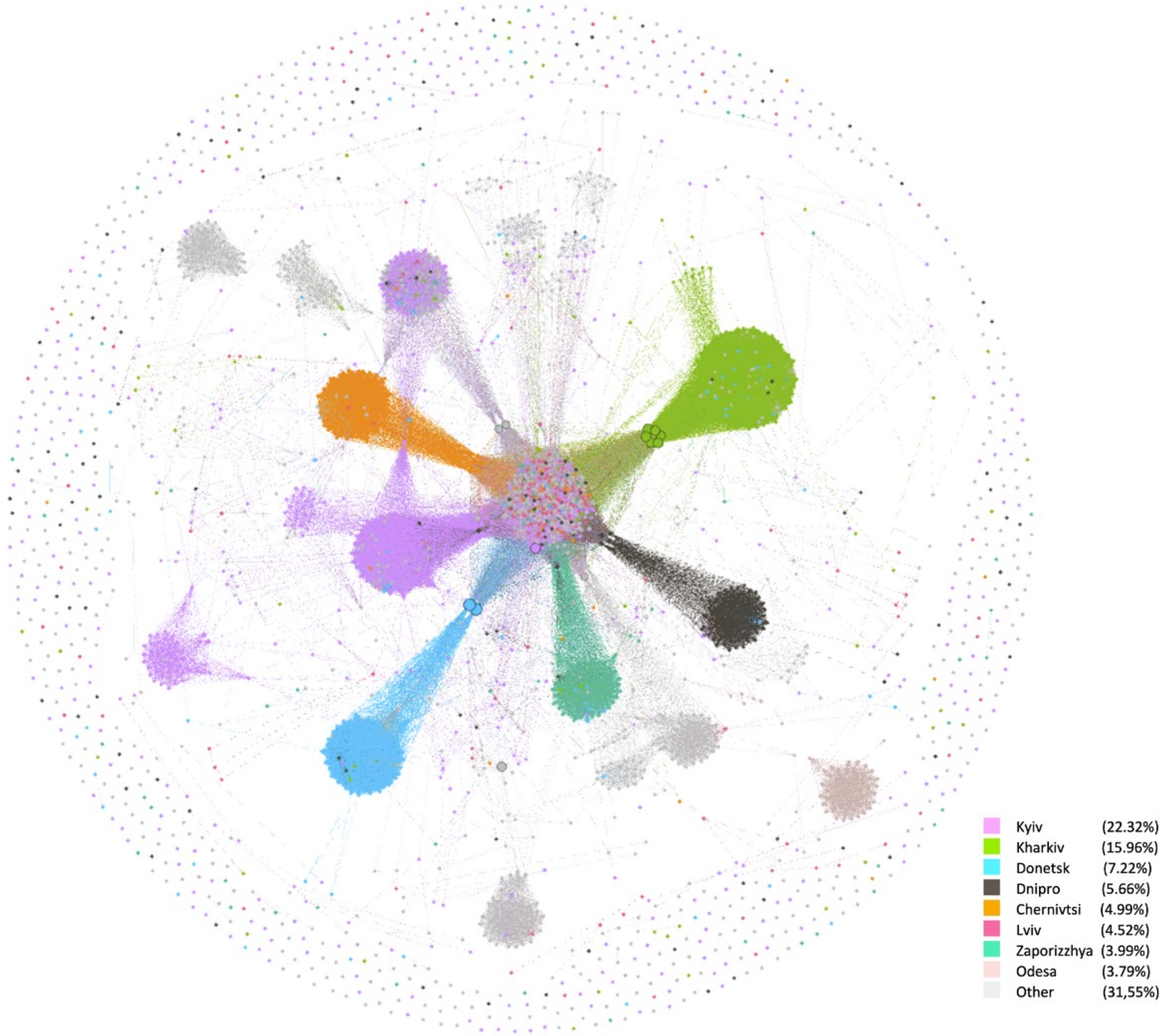
Prosecutors' declarant-organization network visualization.

Reduction to one-mode networks



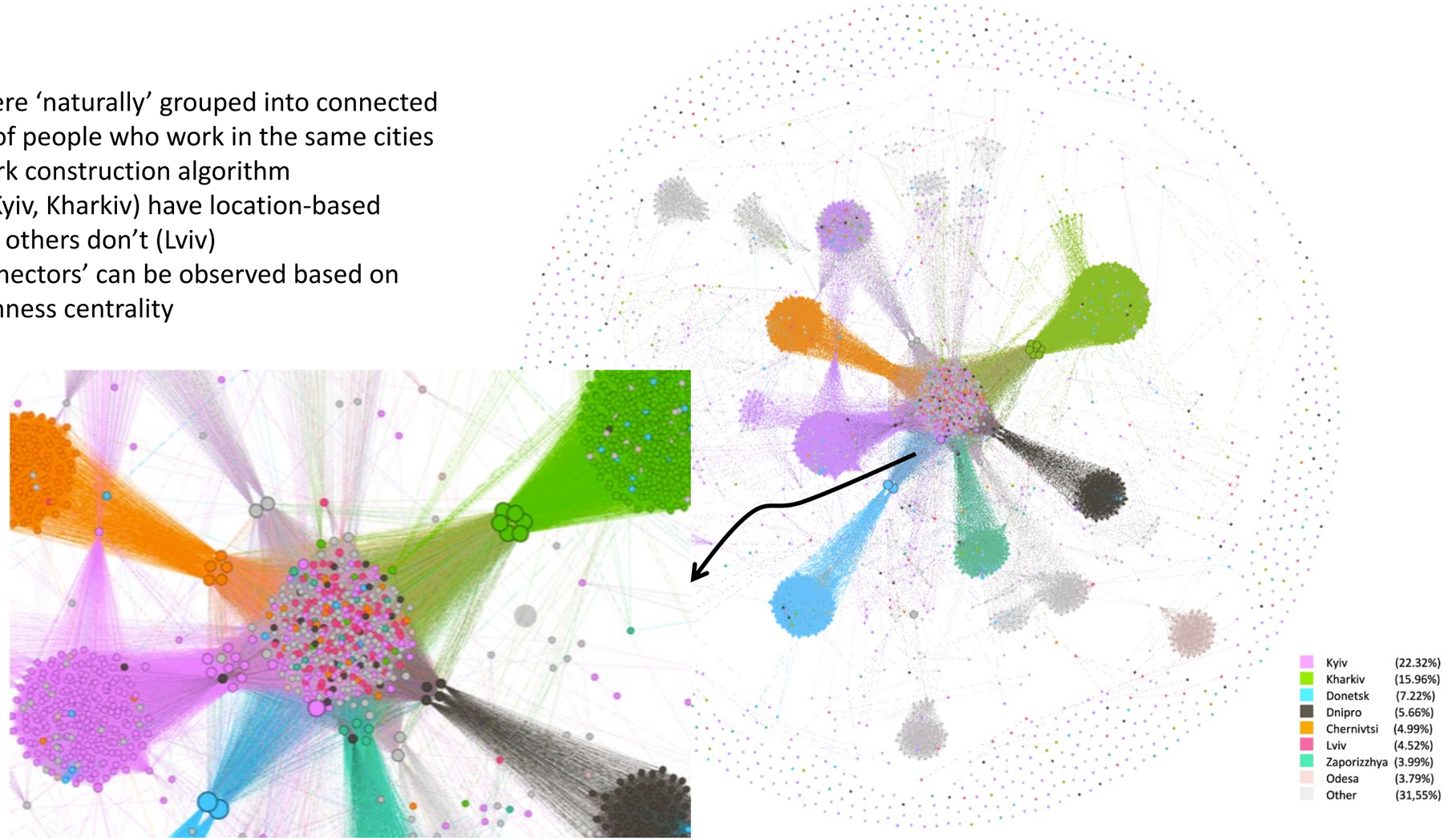
Graphs’ features: location, centrality

- Declarants were ‘naturally’ grouped into connected components of people who work in the same cities by our network construction algorithm
- Some cities (Kyiv, Kharkiv) have location-based communities, others don’t (Lviv)
- Network ‘connectors’ can be observed based on their betweenness centrality



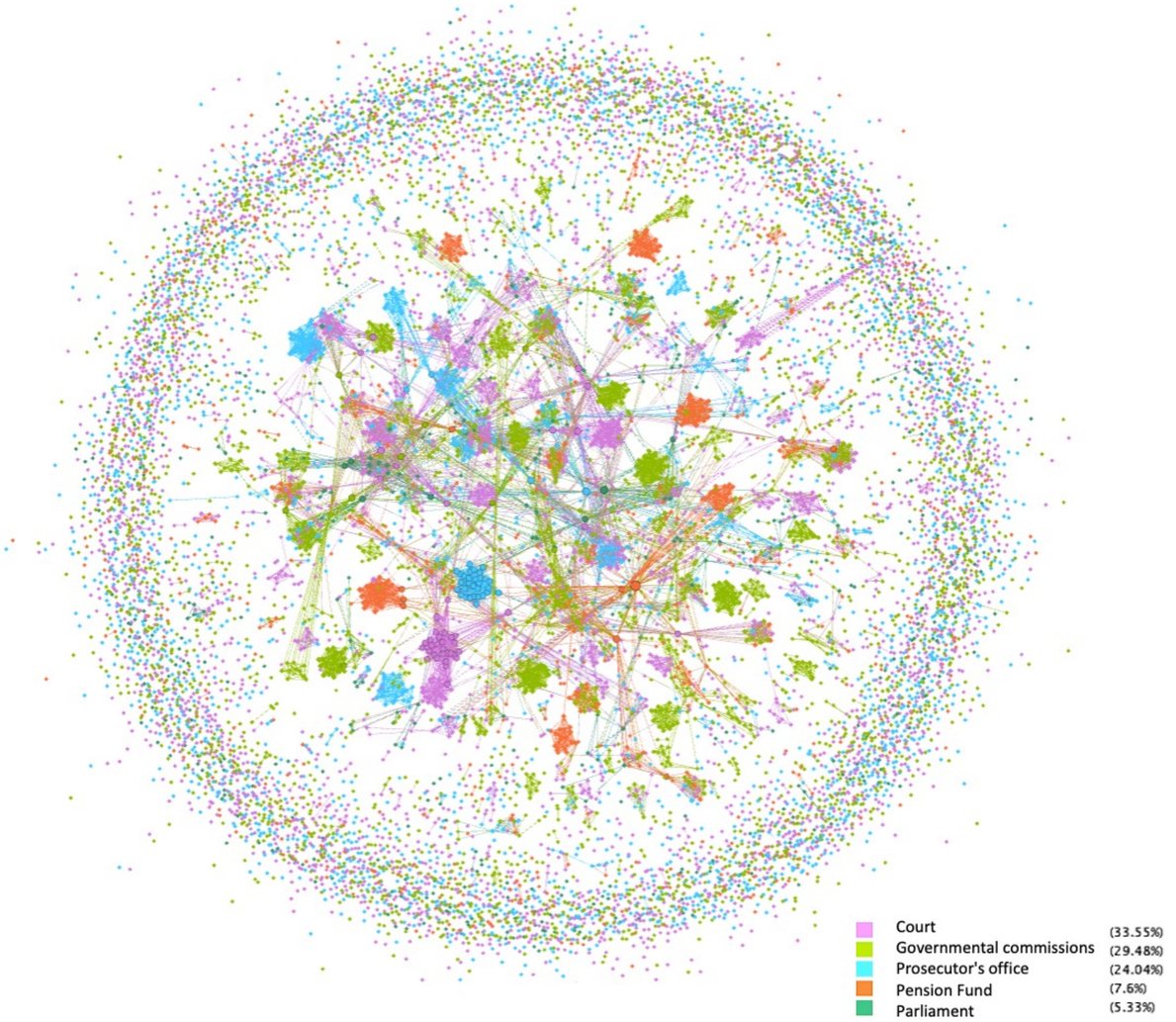
Graphs' features: location, centrality

- Declarants were 'naturally' grouped into connected components of people who work in the same cities by our network construction algorithm
- Some cities (Kyiv, Kharkiv) have location-based communities, others don't (Lviv)
- Network 'connectors' can be observed based on their betweenness centrality



Graphs’ features: offices connectivity

- Large companies and orgs (>30 members) are excluded
- Person-to-person links remain in the graph
- Smaller communities are detectable



Corruption risk profiling with network graphs

- Declarations open dataset is not enough to predict if a governmental worker committed corruption

- Network-based indicators that can be a proxy for corruption risk:
  - Declarant’s level of influence (centrality)
  - Affiliation with high-risk organizations
  - Presence in specific network clusters

- Future work
  - Collaboration with political domain expert
  - Adding more declarant-related features
  - Dynamic graphs
  - Other data sources (e.g., Corruption Register)

Summary

- Analyzed the available data sources and provided motivation to chose declarations.com.ua API
- Identified and extracted features of the declarations that can be used for network modelling
- Suggested and implemented two approaches for network modelling by defining two types of connections between public officials: via company co-ownership and via an organization.
- Built primary two-mode networks and analyzed their key topological characteristics
- Reduced to one-mode networks and showed that they help detect communities and people with high betweenness centrality. Demonstrated communities based on locations, offices, political parties.
- Suggested a set of methods that can be used for future corruption risk profiling with graphs.

Thank you!

Let me know if you have questions or clarifications.

Reviewer's comments discussion

- All in all, the work is nice and easy to follow. Although I appreciate Iryna's careful and transparent analysis, I also think that her thesis could benefit from a **better literature review and engagement with political science ideas** on political influence.

- Her finding could be relevant to predict political influence (instead of corruption) or some other underlying mechanisms of the political process (e.g., party capture). There is a vast scholarship on revolving doors, groups of influence, party capture, networks of elites. This scholarship draws from graph theory and network analysis extensively.

- Most importantly, I expected to see more discussion of particular community detection methods (SNA has many different clustering algorithms) and some SNA-specific metrics (e.g., density, transitivity, assortativity). I also think that this empirical work could benefit from running some models (QAP correlations or ERGM models).

- Despite some limitations, I believe that this is solid work. I especially appreciate the citation of Nicolas Christakis and Iryna's final sentence that "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". I share her sentiment and believe that this is a very nice conclusion for any person who aspires to be a researcher.

APPENDIX

Related works

- OGD application development : “Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives” by Barbara Ubaldi (Ubaldi, 2013).
- 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)
- 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).
- 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.
- 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.
- “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 behaviour of a procurement network.

Open Data Principles

1. Open by Default

2. Timely and Comprehensive

3. Accessible and Usable

4. Comparable and Interoperable

5. For Improved Governance and Citizen Engagement

6. For Inclusive Development and Innovation

Declaration example

https://declarations.com.ua/declaration/nacp_294babb4-0db7-4692-bd0f-ad9f2f031ba6

Network metrics used in the work

- **Average Degree** is the average number of edges per node in the graph. It is relatively straightforward to calculate.
  Total Edges / Total Nodes = Average Degree

- **Network density** - the number of existing relationships relative to the possible number. Dense networks are more important for control and sanctioning than for information. Dense networks tend to generate a lot of redundant information.

- **Modularity coefficient** is a measure of the structure of networks or graphs which measures the strength of division of a network into modules (also called groups, clusters or communities).
  Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000

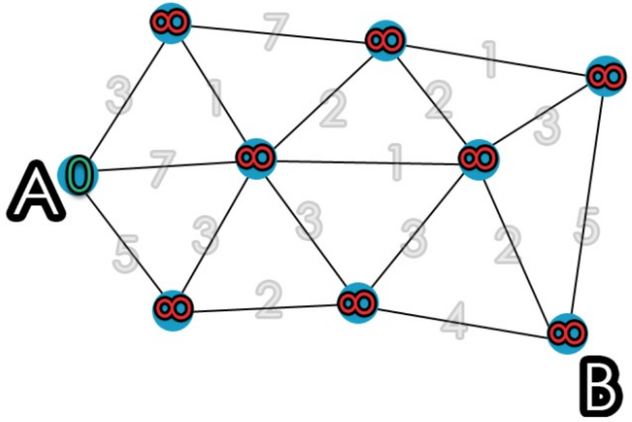
- **Connected components**
  Robert Tarjan, Depth-First Search and Linear Graph Algorithms, in SIAM Journal on Computing 1 (2): 146–160 (1972)

- **Betweenness centrality** captures how much a given node (hereby denoted u) is in-between others. This metric is measured with the number of shortest paths (between any couple of nodes in the graphs) that passes through the target node u (denoted σσv,w(u)).

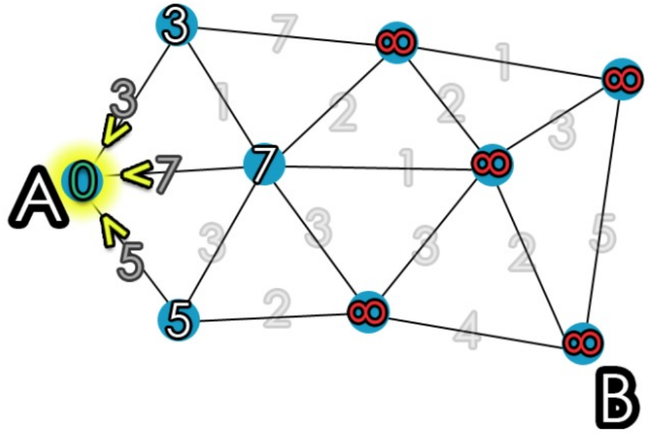
- Community detection in network graphs (Blondel et al., 2008)

Dijkstra's algorithm for shortest path

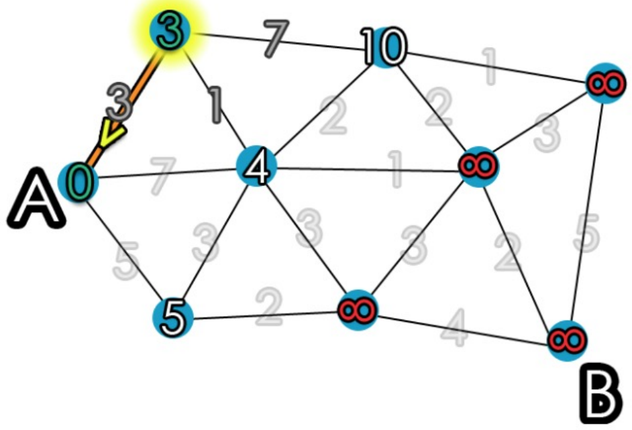
1 Initialize distances according to the algorithm.



2 Pick first node and calculate distances to adjacent nodes.



3 Pick next node with minimal distance; repeat adjacent node distance calculations.



4 Final result of shortest-path tree.