The Bibi Graph

A social network analysis of leaders' autobiography via NER, language models, and graph algorithms

Gal Engelberg, January 2023

Motivation

An autobiography represents a first-person narrative that covers the writer's life from their own perspective.

Leaders' autobiographies usually express their agenda, and a narrative that the leaders want to be remembered according to.

Could this narrative be quantified? Is there a way to encode the textual information in a way that will uncover global insights about the leader's perspective?

The proposed approach & questions

Approach:

Extracting **named entities** out of the autobiography to the form of a **social network** and analyze the resulting graph via social network analysis algorithms. This will uncover insights such as:

Questions:

What are the **cardinal persons** according to the writer?

What are the **cardinal events** according to the writer?

Can we discover **meaningful communities** in the graph?

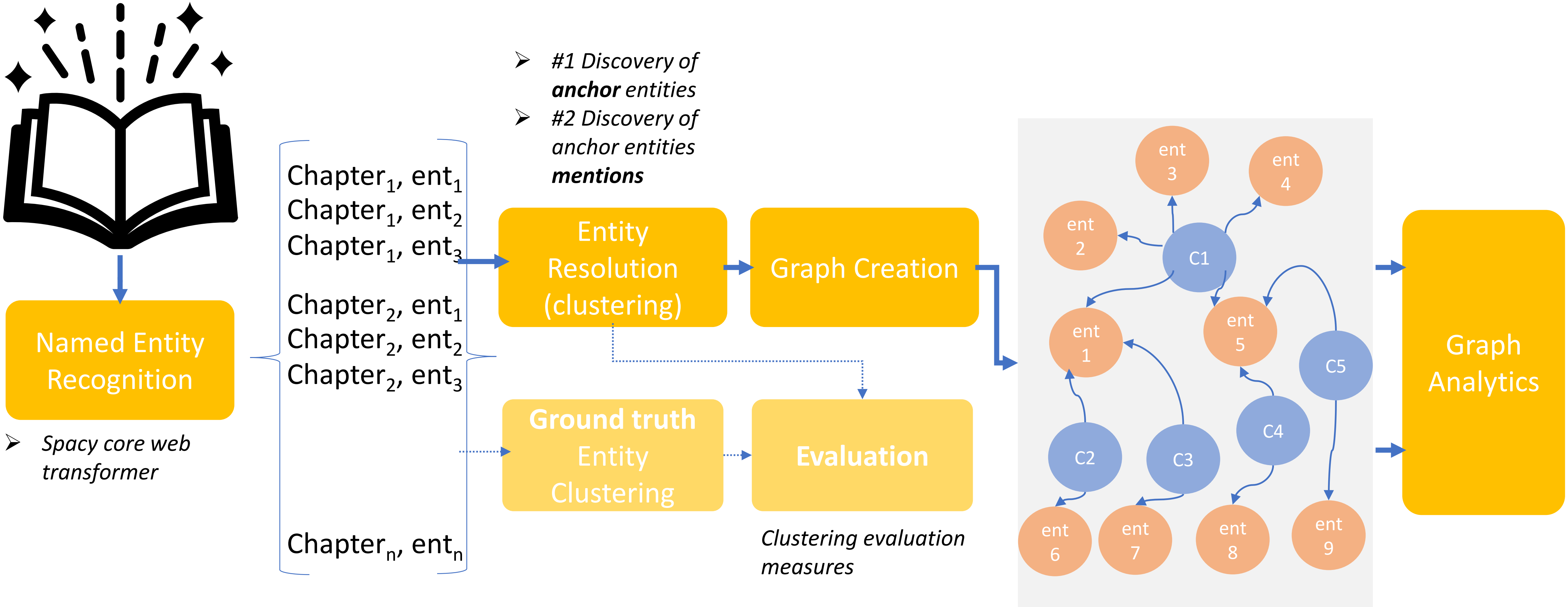
Challenges

- ➢ How to recognize named entities within the autobiography?
- ➢ How to perform entity resolution?

  “Barack Obama”, “Obama” => “Barack Obama”

- ➢ How to represent the extracted entities in a graphical structure?
- ➢ How to analyze the graph to uncover insights from the book?

Methodology



Named Entity Recognition

- *Spacy core web transformer*

Chapter$_1$, ent$_1$
Chapter$_1$, ent$_2$
Chapter$_1$, ent$_3$

Chapter$_2$, ent$_1$
Chapter$_2$, ent$_2$
Chapter$_2$, ent$_3$

Chapter$_n$, ent$_n$

- *#1 Discovery of* ***anchor*** *entities*
- *#2 Discovery of anchor entities* ***mentions***

Entity Resolution (clustering)

Graph Creation

Ground truth Entity Clustering

Evaluation

Clustering evaluation measures

Graph Analytics

Data source

88 chapters

250K Token instances

32K Tokens

8.4K named entities

1,700 persons
635 unique persons

465 Events
66 unique events

Named entity recognition

- Language model: Spacy English core web transformer (Roberta-based) spacy/en_core_web_trf
- NER was trained via OntoNote project
  - **Multiple genres** news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows
  - **Three languages** (English, Chinese, and Arabic)
  - **Mapping to shallow semantics** (ontology and coreference)

Entity resolution (clustering)

Events

- Events named entity represented as word count vectors
- We run a DBSCAN clustering algorithm with the following parameters:
  - Maximum distance: 0.3
  - Similarity measure: cosine similarity

Entity resolution (clustering)

Persons

- Persons named entity represented as 3-gram character level count vectors
- We performed two steps:
  - Step 1: discovery of anchor persons

    Clustering of named entities which are mentioned with their full names
    (DBSCAN, maximal distance = 0.25, Similarity measure: cosine similarity)
  - Step 2: discovery of anchor persons' mentions in the text

```
foreach named_entity within a chapter:
  a = anchor_with_min_distance(named_entity, chapter_level_anchors) //cosine similarity
  if a.distnace > threshold: //threshold = 0.5
    map (named_entity, a)
  else:
    b = anchor_with_min_distance(named_entity, book_level_anchors) //cosine similarity
    if a.distnace > threshold: //threshold = 0.5
      map (named_entity, a)
```

Evaluation measures

Measure	Description
Adjusted Mutual Information Score	The Mutual Information is a measure of the similarity between two labels of the same data. Adjusted Mutual Information (AMI) is an adjustment of the Mutual Information (MI) score to account for chance. It accounts for the fact that the MI is generally higher for two clusterings with a larger number of clusters, regardless of whether there is actually more information shared.
Adjusted Rand Score	The Rand Index computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. The adjusted Rand index is thus ensured to have a value close to 0.0 for random labeling independently of the number of clusters and samples and exactly 1.0 when the clusterings are identical (up to a permutation). The adjusted Rand index is bounded below by -0.5 for especially discordant clusterings.
Completeness Score	A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster.
Homogeneity Score	A clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class.
V Measure Score	The V-measure is the harmonic mean between homogeneity and completeness

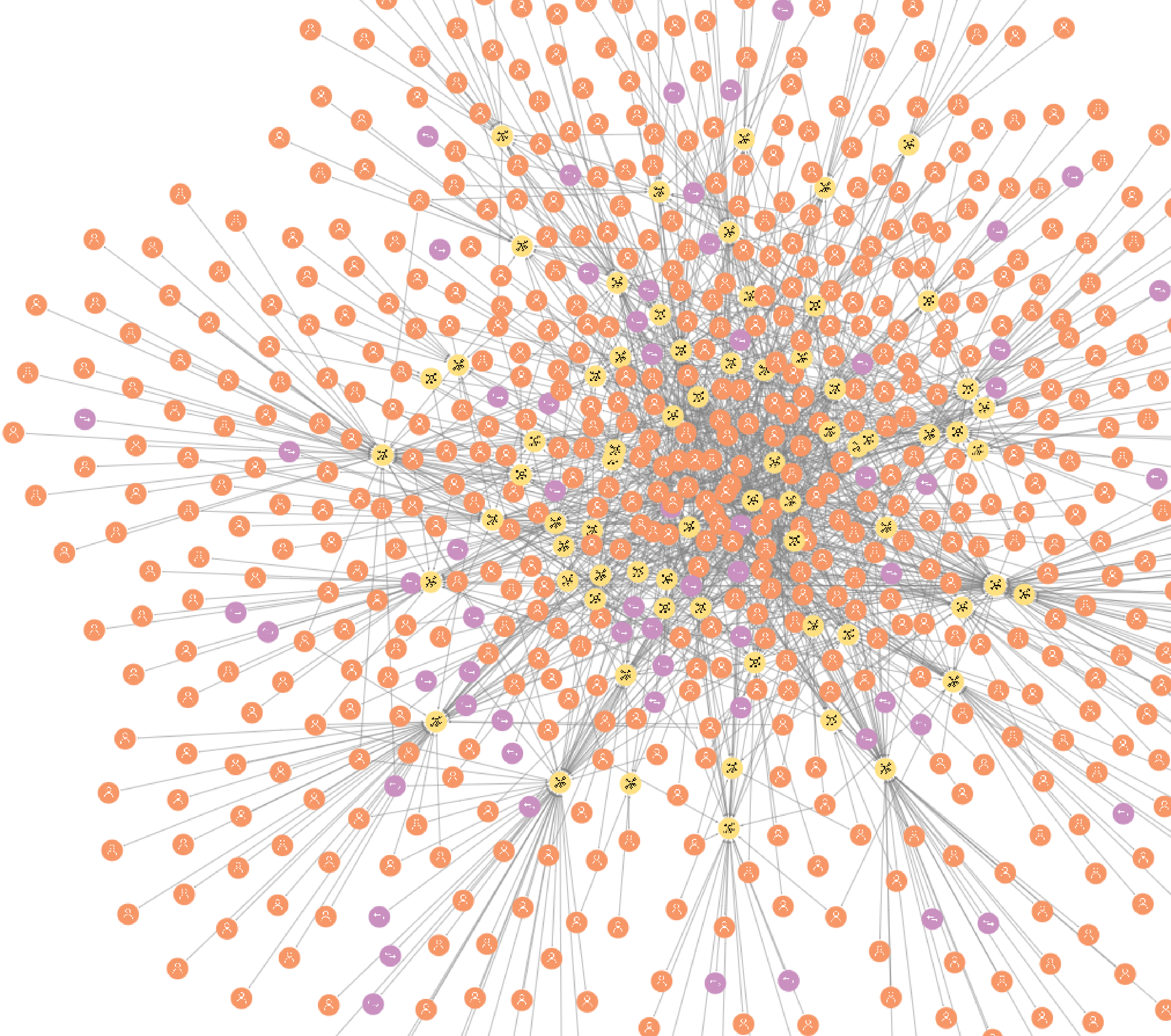
Evaluation results

- Evaluation was performed by comparing the entity resolution resulted partitions with the ground truth partitions

Measure	Events clustering	Persons clustering
Adjusted Mutual Information Score	0.96	0.92
Adjusted Rand Score	0.92	0.82
Completeness Score	0.98	0.96
Homogeneity Score	0.97	0.95
V Measure Score	0.97	0.95

Results are high across all measures and entity types

Graph Structure and Statistics



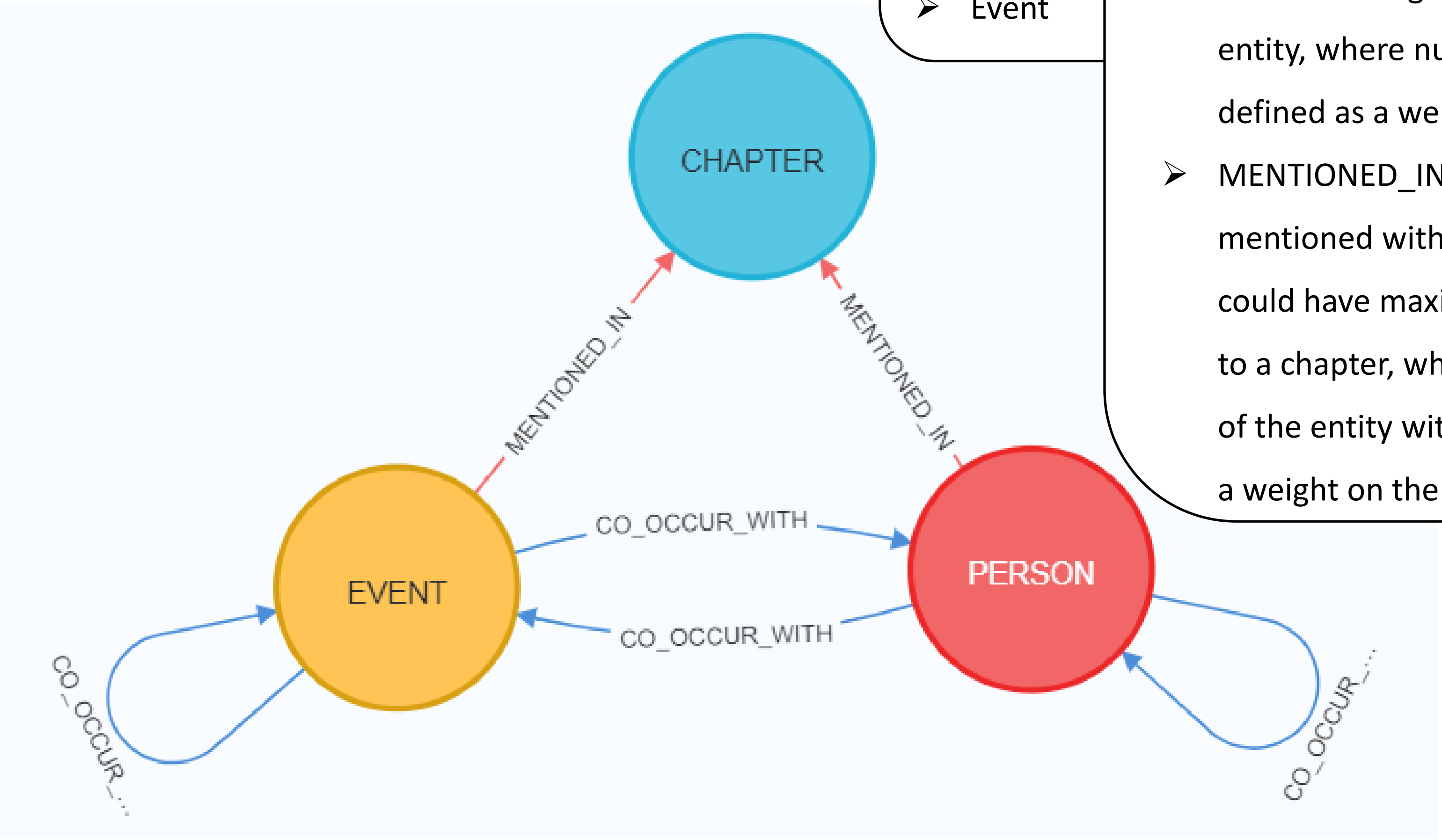
Graph structure

Nodes

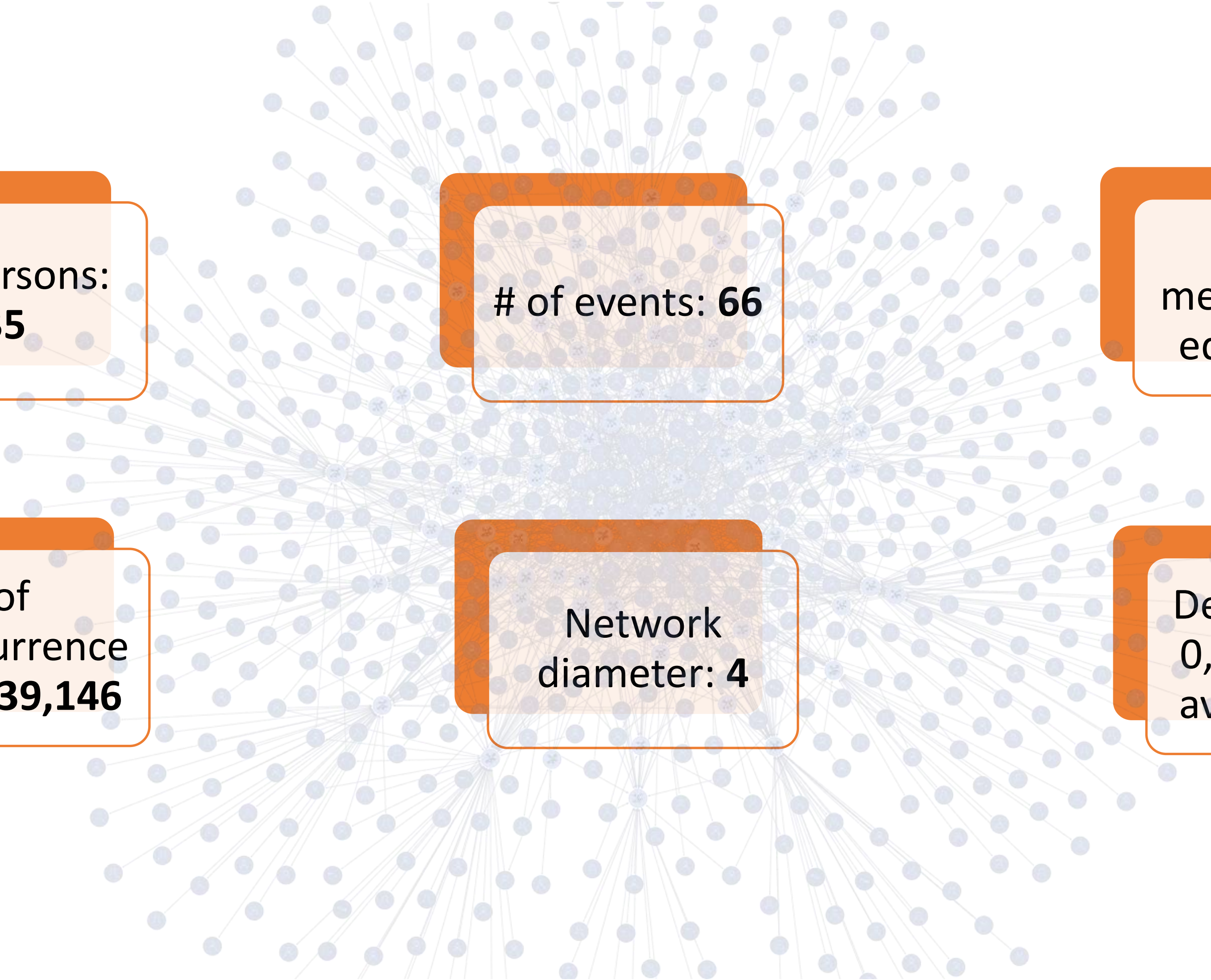
- Chapter
- Person
- Event

Relations

- CO_OCCUR_WITH: Co-occurrence of two entities in a chapter. An entity could have maximum single outgoing edge to another entity, where number of joint chapters is defined as a weight on the edge.
- MENTIONED_IN: Whether an entity was mentioned within the chapter. An entity could have maximum single outgoing edge to a chapter, where the amount of mentions of the entity within the chapter is defined as a weight on the edge.



Graph statistics



# of persons: **635**

# of events: **66**

# of mentioned_in edges: **1590**

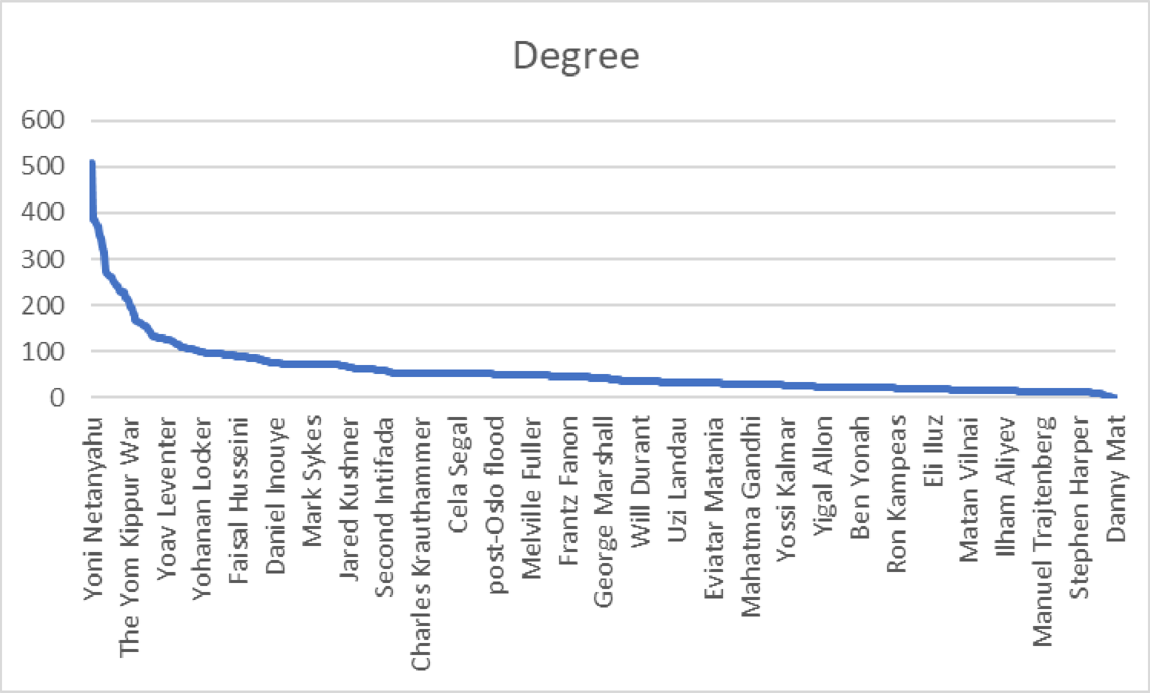
# of co_occurrence edges: **39,146**

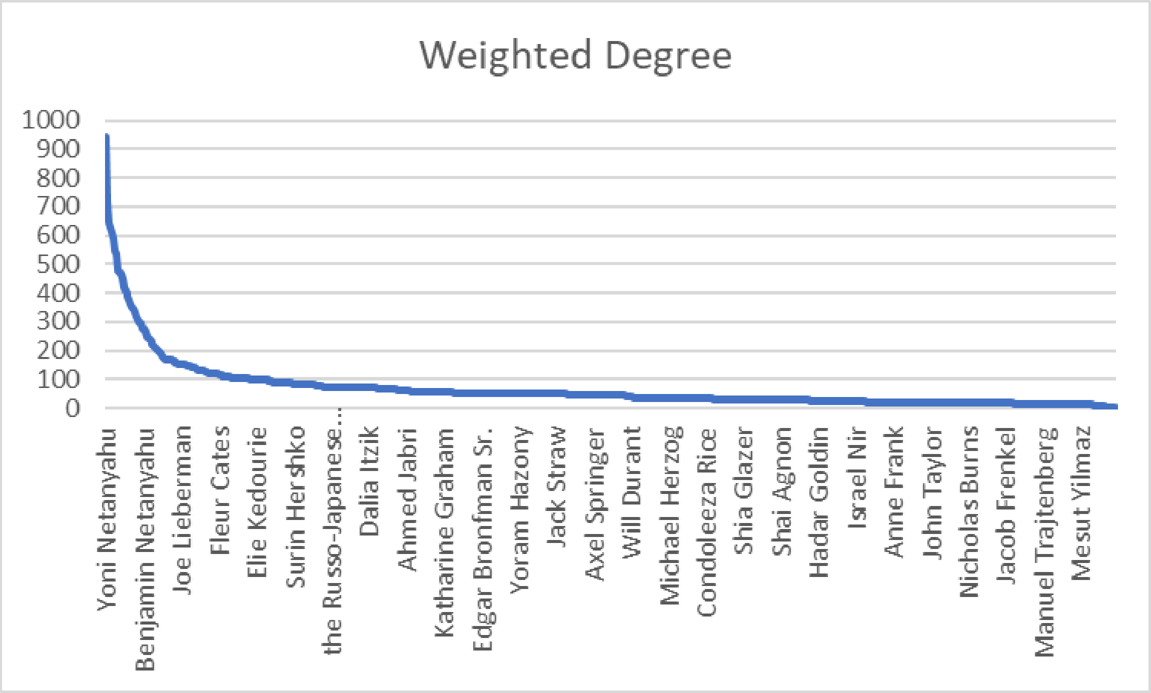
Network diameter: **4**

Degree (min = 0, max = 509, average = 55)

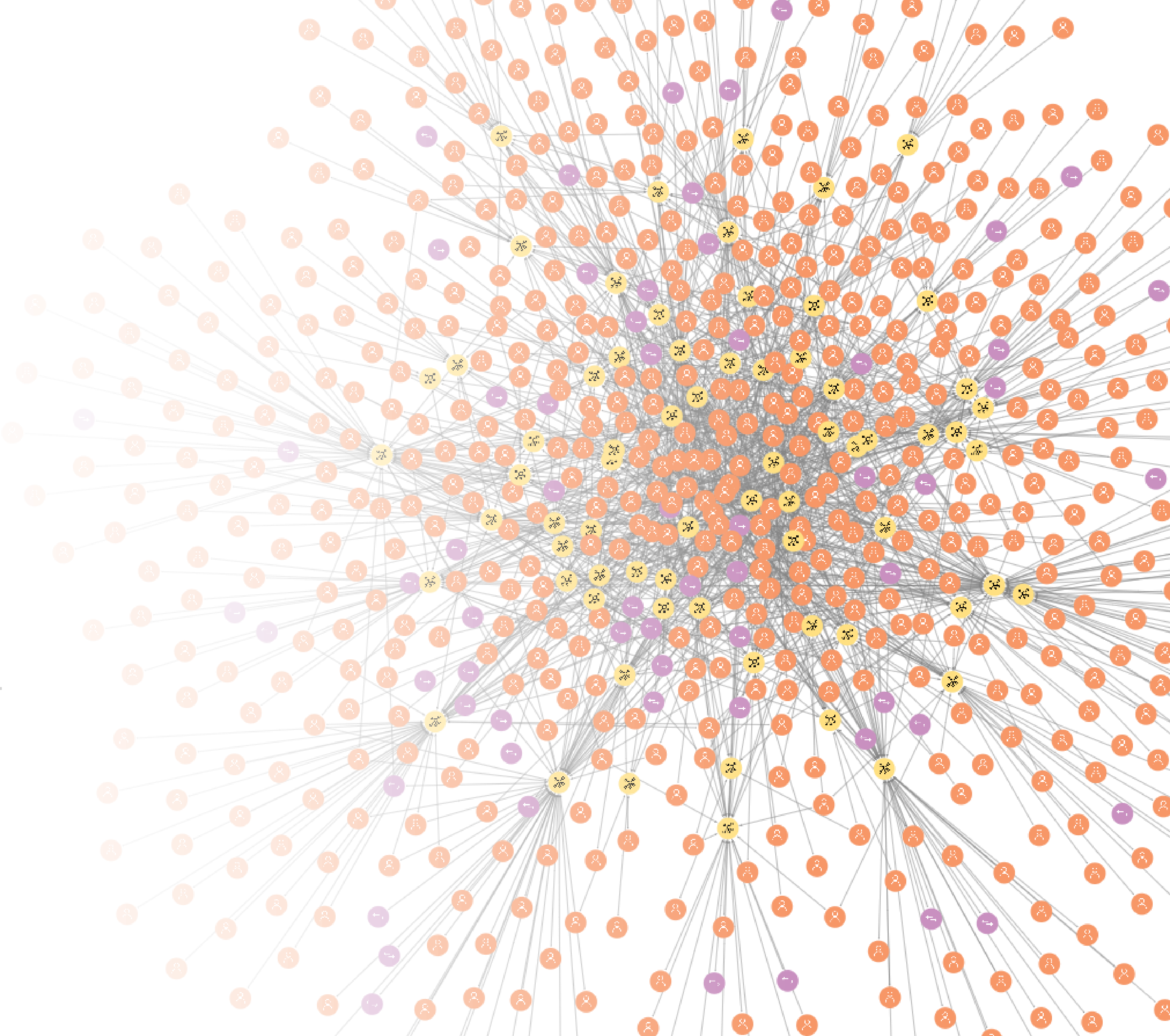
Entities degree

based on CO_OCCUR_WITH edges





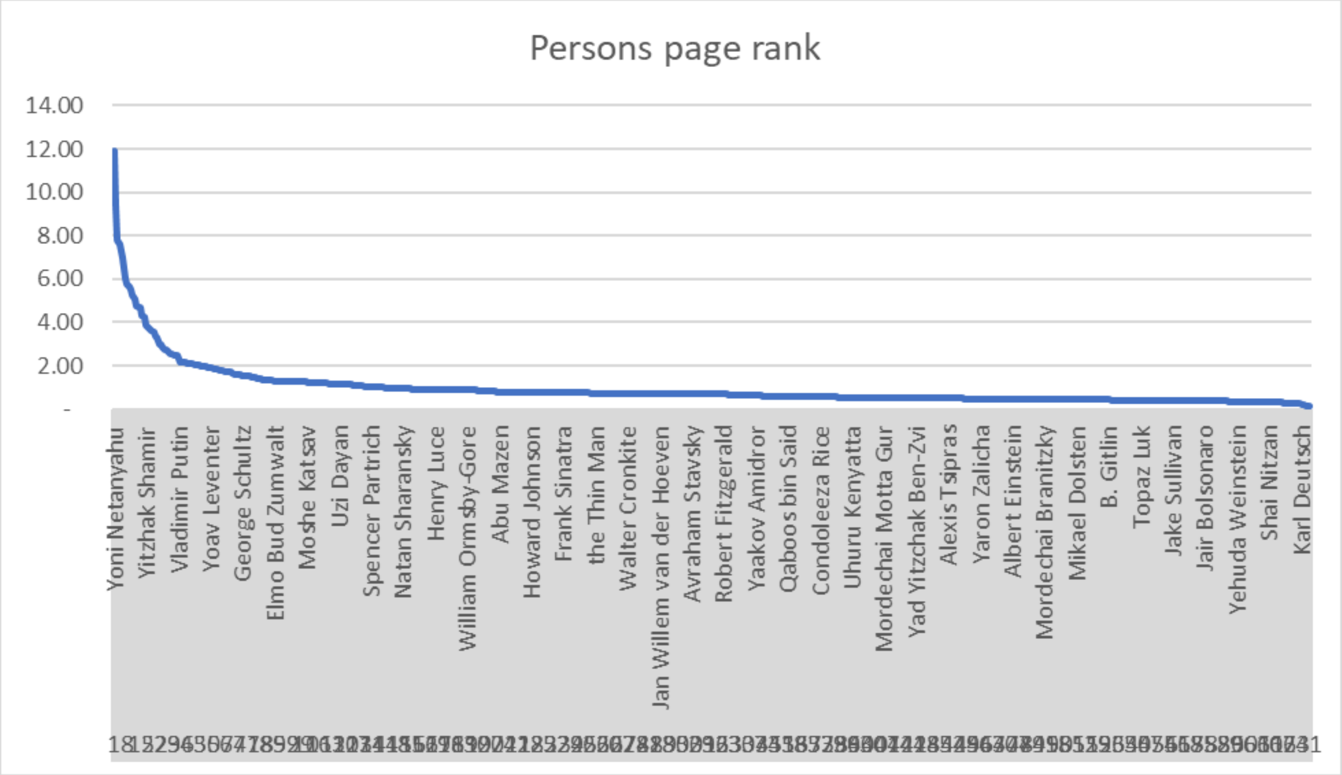
What are the cardinal persons according to the writer?



Cardinal persons

Page rank (based on CO_OCCUR_WITH edges, weighted)

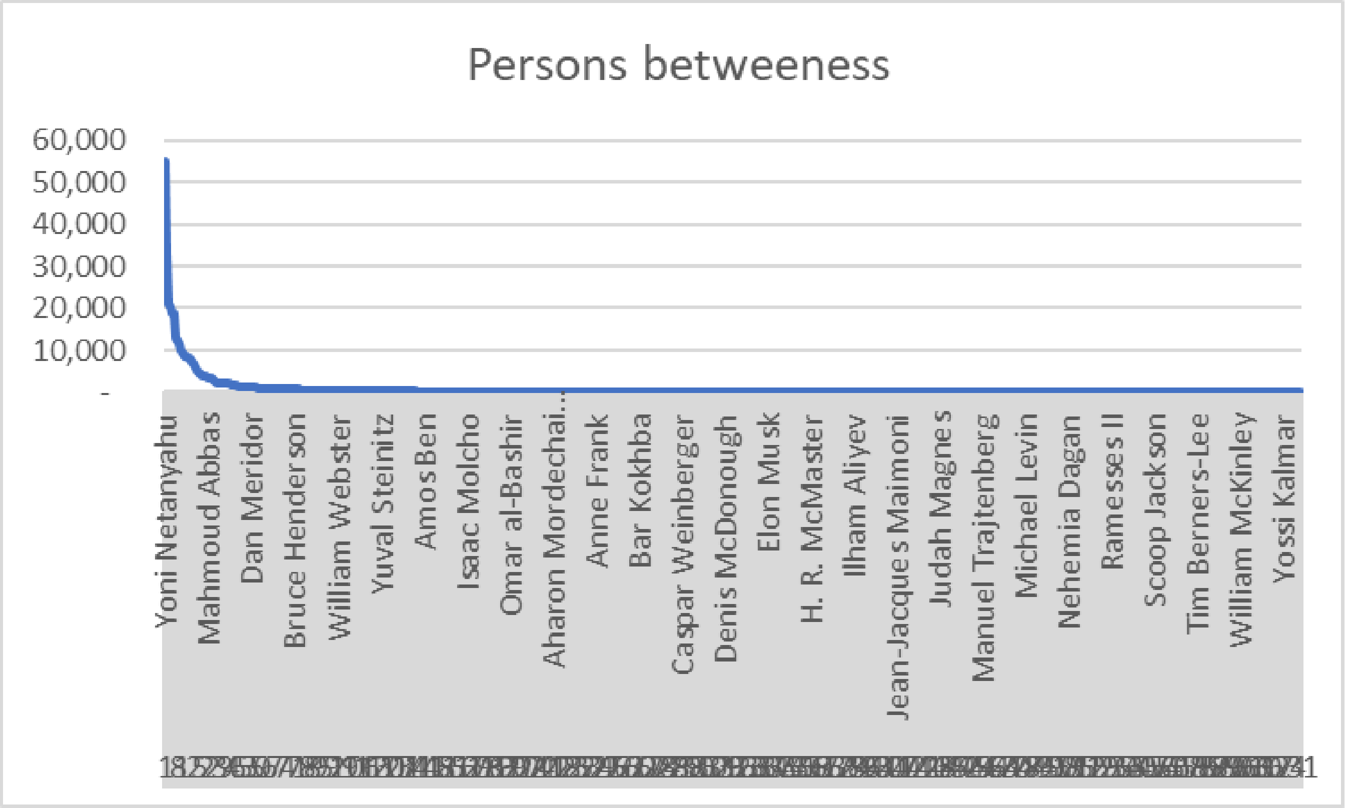
#	name	score
1	Yoni Netanyahu	11.93
2	Sara Netanyahu	9.36
3	SHIMON PERES	7.81
4	Yitzhak Rabin	7.59
5	Ariel Sharon	7.26
6	Benzion Netanyahu	6.95
7	Barack Obama	6.05
8	Zeev Jabotinsky	5.75
9	Yair Lapid	5.64
10	Iddo Netanyahu	5.50
11	Ron Dermer	5.24
12	Yasser Arafat	5.06
13	Avner Schur	4.74
14	Menachem Begin	4.72
15	Ehud Barak	4.71
16	Ben Gurion.1	4.30
17	Bill Clinton	4.23
18	Yitzhak Shamir	3.87
19	Ehud Olmert	3.80
20	Benjamin Netanyahu	3.67
21	Mahmoud Abbas	3.60
22	Ronald Reagan	3.54
23	Moshe Arens	3.37
24	Melania Trump	3.31
25	Theodore Herzl	3.01



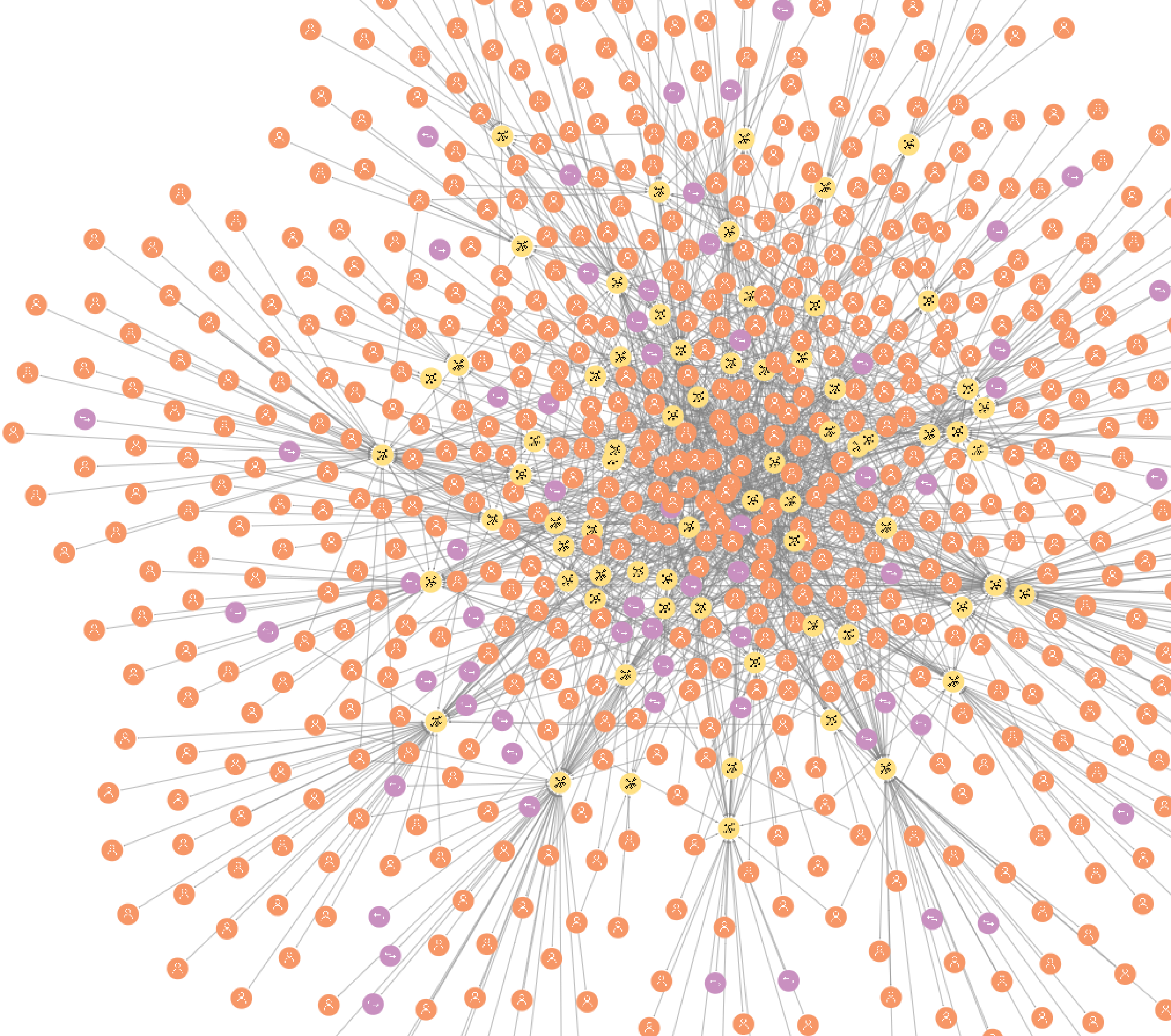
Cardinal persons

Betweenness (based on CO_OCCUR_WITH edges , weighted)

#	name	score
1	Yoni Netanyahu	55,003
2	Sara Netanyahu	32,617
3	SHIMON PERES	20,744
4	Ariel Sharon	20,347
5	Benzion Netanyahu	18,772
6	Yitzhak Rabin	18,631
7	Barack Obama	12,613
8	Iddo Netanyahu	12,126
9	Zeev Jabotinsky	11,420
10	Benjamin Netanyahu	9,725
11	Ehud Barak	9,555
12	Yasser Arafat	8,554
13	Menachem Begin	8,332
14	Yair Lapid	8,169
15	Ron Dermer	8,078
16	Bill Clinton	7,420
17	Ben Gurion.1	6,716
18	Avner Schur	5,989
19	Saddam Hussein	5,026
20	Ronald Reagan	4,694
21	Yitzhak Shamir	4,111
22	Winston Churchill	3,995
23	Melania Trump	3,968
24	Vladimir Putin	3,950
25	Mahmoud Abbas	3,564



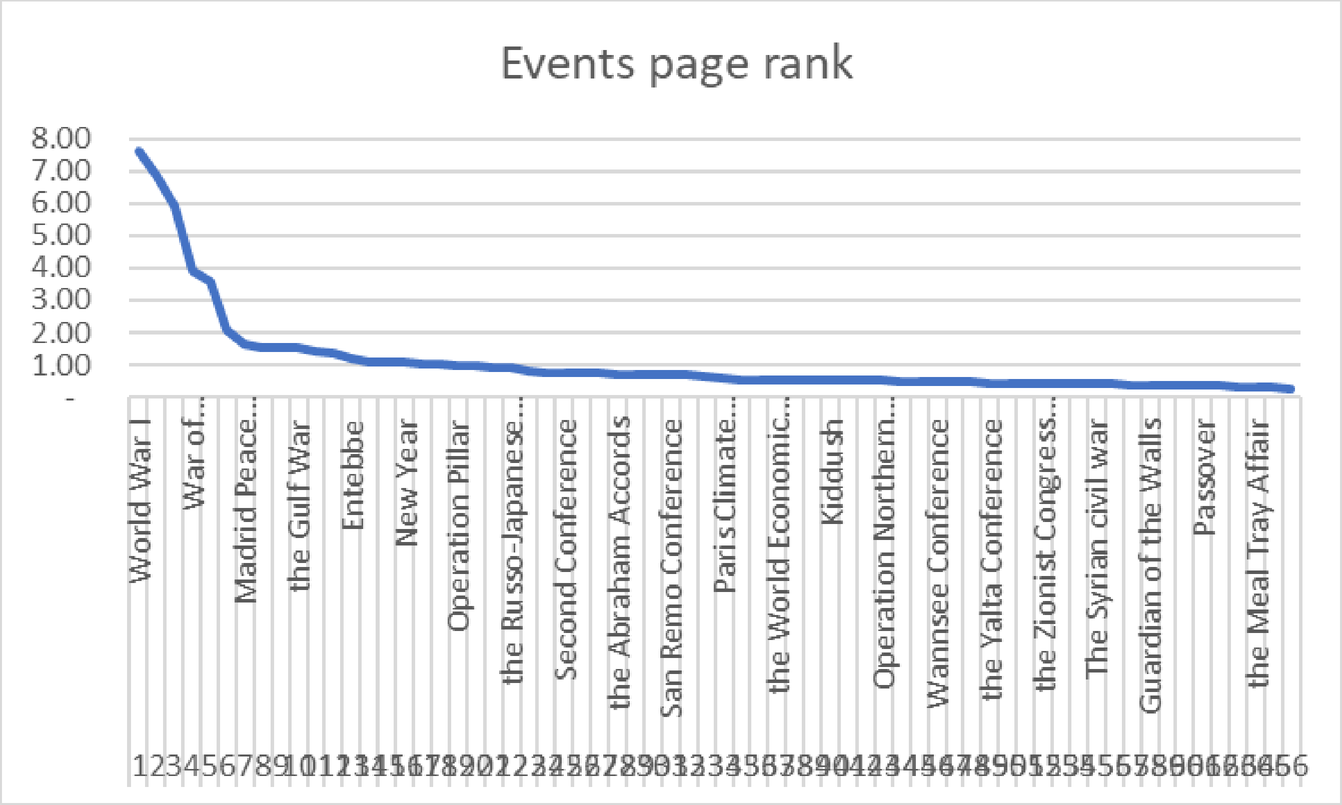
What are the cardinal events according to the writer?



Cardinal events

Page rank (based on CO_OCCUR_WITH edges , weighted)

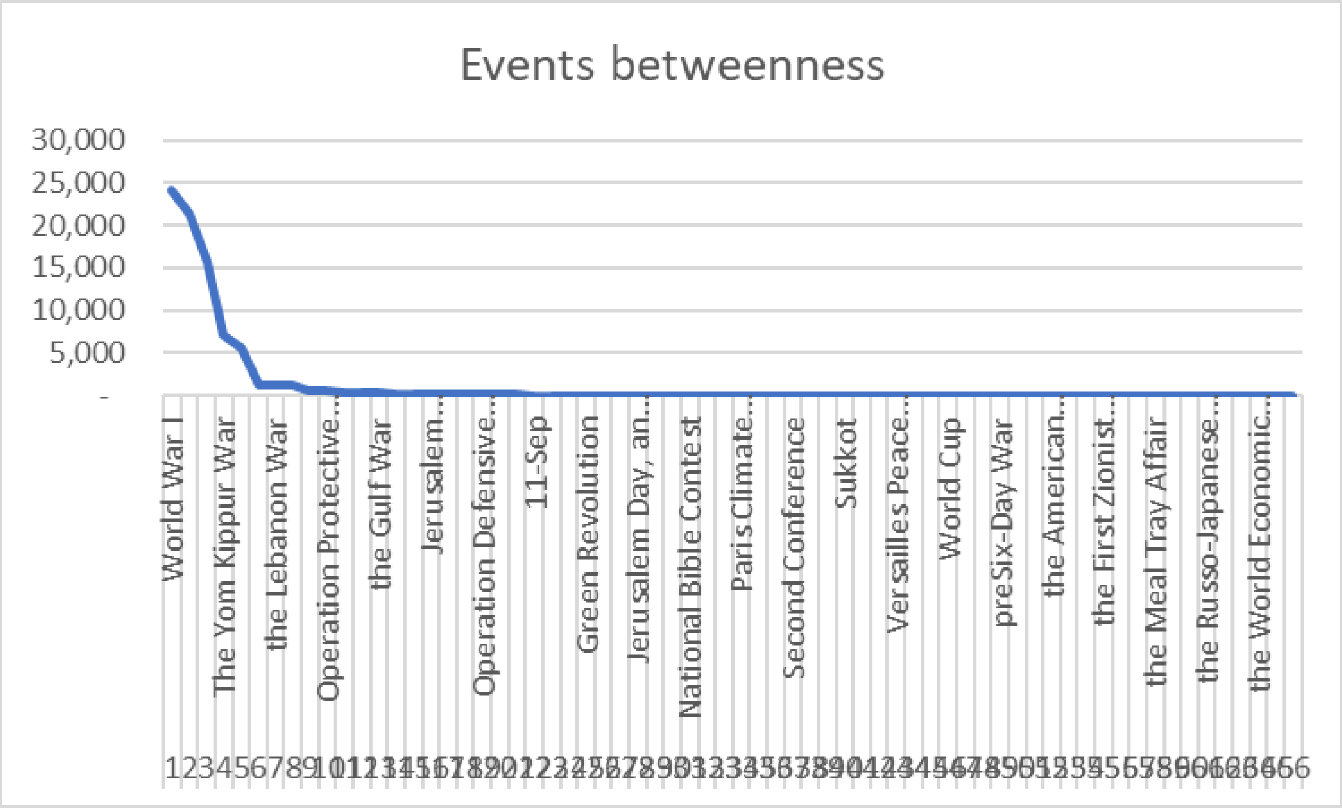
#	name	score
1	World War I	7.60
2	Holocaust	6.86
3	the Six-Day War	5.88
4	War of Independence (1948	3.91
5	The Yom Kippur War	3.57
6	the Lebanon War	2.06
7	Madrid Peace Conference	1.61
8	Suez	1.53
9	the Arab Spring	1.53
10	the Gulf War	1.50
11	Jerusalem Conference on International Terrorism	1.39
12	the Washington Conference	1.34
13	Entebbe	1.22
14	Operation Defensive Shield	1.10
15	Operation Protective Edge.9	1.09
16	New Year	1.06
17	Oslo	1.05
18	Second Intifada	1.03
19	Operation Pillar	0.97
20	the Arlosoroff Affair	0.97



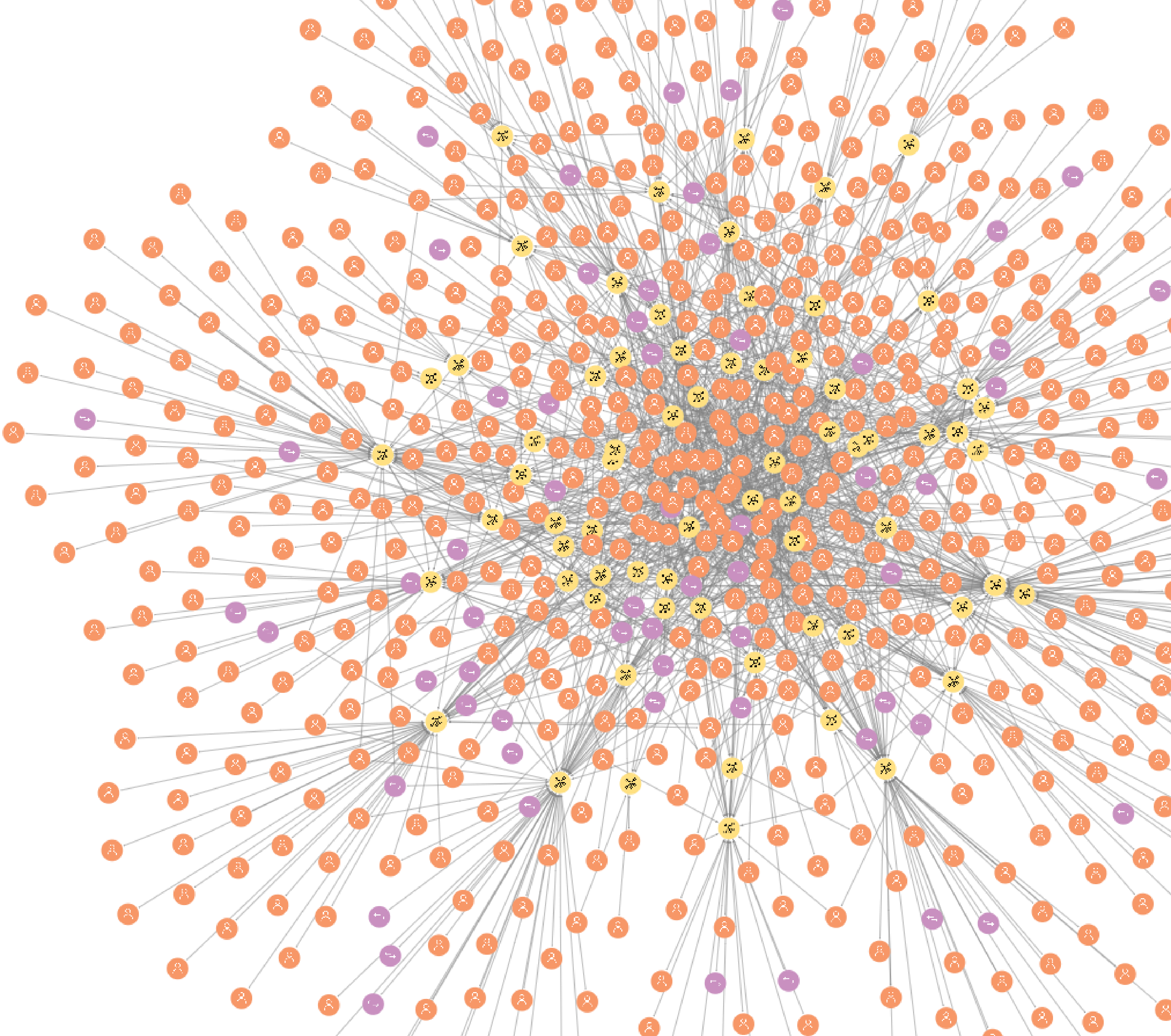
Cardinal events

Betweenness(based on CO_OCCUR_WITH edges , weighted)

#	name	score
1	World War I	24,197
2	Holocaust	21,477
3	the Six-Day War	15,711
4	The Yom Kippur War	7,063
5	War of Independence (1948	5,586
6	the Arab Spring	1,208
7	the Lebanon War	1,198
8	Suez	1,070
9	Second Intifada	635
10	Operation Protective Edge.9	562
11	Madrid Peace Conference	403
12	Entebbe	385
13	the Gulf War	385
14	the Washington Conference	214
15	the Vietnam War	211
16	Jerusalem Conference on International Terrorism	208
17	Operation Pillar	199
18	New Year	122
19	Operation Defensive Shield	98
20	the Arlosoroff Affair	84



Can we discover meaningful communities in the graph?



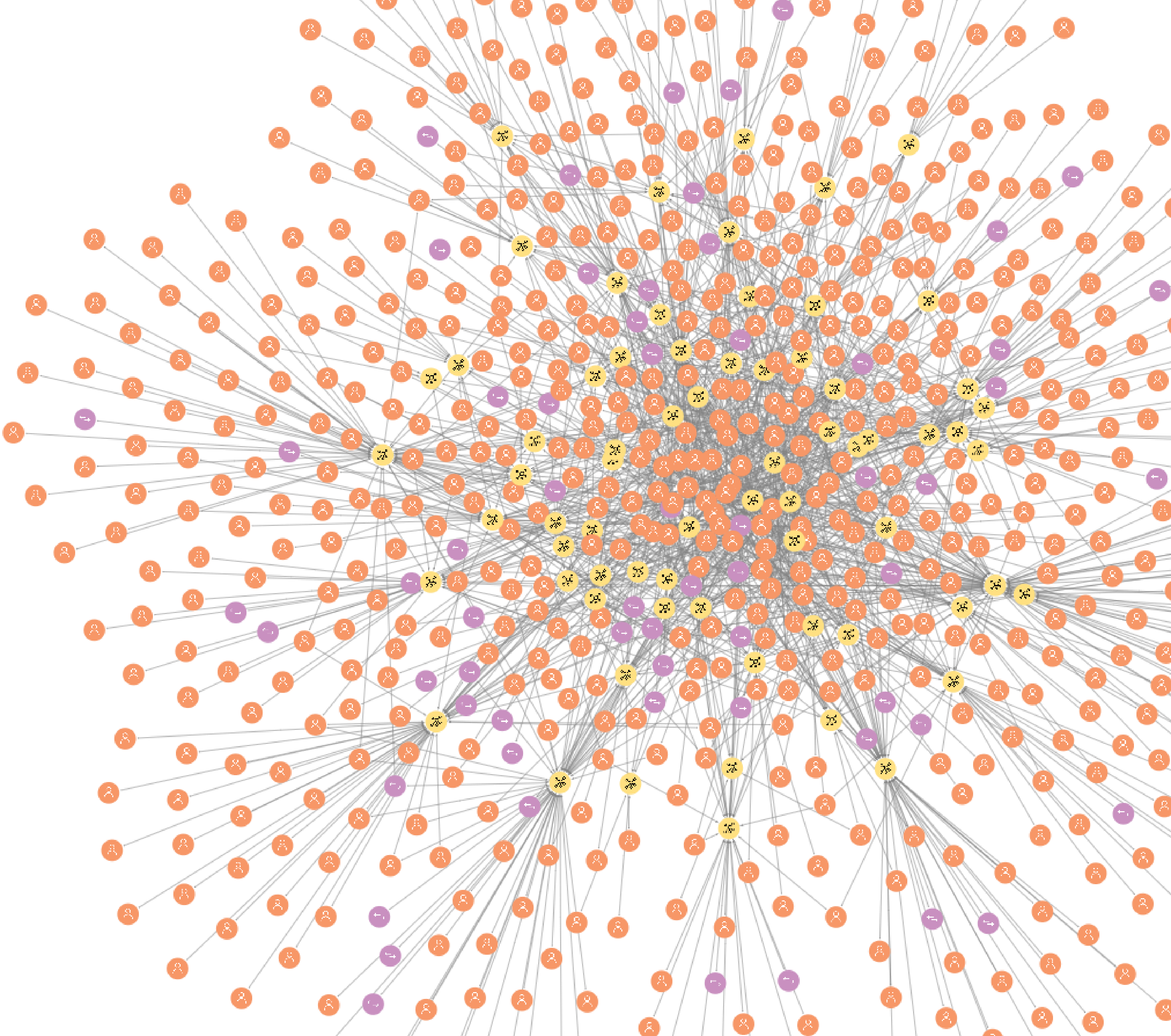
Community Detection

Louvain (based on CO_OCCUR_WITH edges)

ID	# of entities	% of total	Avg. page rank	Top 10 entities In terms of Page-Rank
51	63	9.0%	1.59	Zeev Jabotinsky,Ben Gurion.1,War of Independence (1948,Yitzhak Shamir,Ronald Reagan,Moshe Arens,Theodore Herzl,Jonathan Pollard,Vladimir Hitler,Jonathan Liss
74	67	9.6%	1.11	World War I,Menachem Begin,Saddam Hussein,Margaret Thatcher,George Will,the Lebanon War,Paul Johnson,the Gulf War,Uzi Yairi,Jerusalem Conference on International Terrorism
22	37	5.3%	0.97	Yoav Leventer,Miki (Miriam) Weissman,George Schultz,the Lubavitcher Rebbe,the Arab Spring,Fleur Cates,Jeane Kirkpatrick,the Washington Conference,Vladimir Bukovsky,Amir Ofer
466	101	14.4%	0.91	Ariel Sharon,Yasser Arafat,Ehud Barak,Bill Clinton,Hussein Agha,Hosni Mubarak,Hafez Assad,Elizabeth Gentieu,Madrid Peace Conference,Benny Begin
372	201	28.7%	0.90	Sara Netanyahu,Benzion Netanyahu,Holocaust,Barack Obama,Yair Lapid,Ron Dermer,Avner Schur,Ehud Olmert,Mahmoud Abbas,Melania Trump
18	147	21.0%	0.83	Yoni Netanyahu,SHIMON PERES,Yitzhak Rabin,the Six-Day War,Iddo Netanyahu,The Yom Kippur War,Idi Amin,Motta Gur,Levi Eshkol,Suez
50	82	11.7%	0.81	Benjamin Netanyahu,Sarah Mileikowsky-Netanyahu,Joseph Klausner,Herod the Great,Itzik Molcho,Jesus of Nazareth,Abraham Marcus,Dan Meridor,Saeb Erekat,Ami Ayalon
197	1	0.1%	0.15	Yitzhak Navon
430	1	0.1%	0.15	Danny Mat
676	1	0.1%	0.15	Meir Har Zion

- ➢ The graph has a reasonable modularity of **0.38**
- ➢ Communities have reasonable distribution in terms of number of entities and average page-rank
- ➢ There are three communities of a single entity (these entities does not have co occurrence with other entities)
- ➢ It seems that communities are arranged by epochs. For example, Zeev Jabotinsky and Ben Gurion, Ariel Sharon and Yasser Arafat.

Further Analytics



Most “intensive” chapters

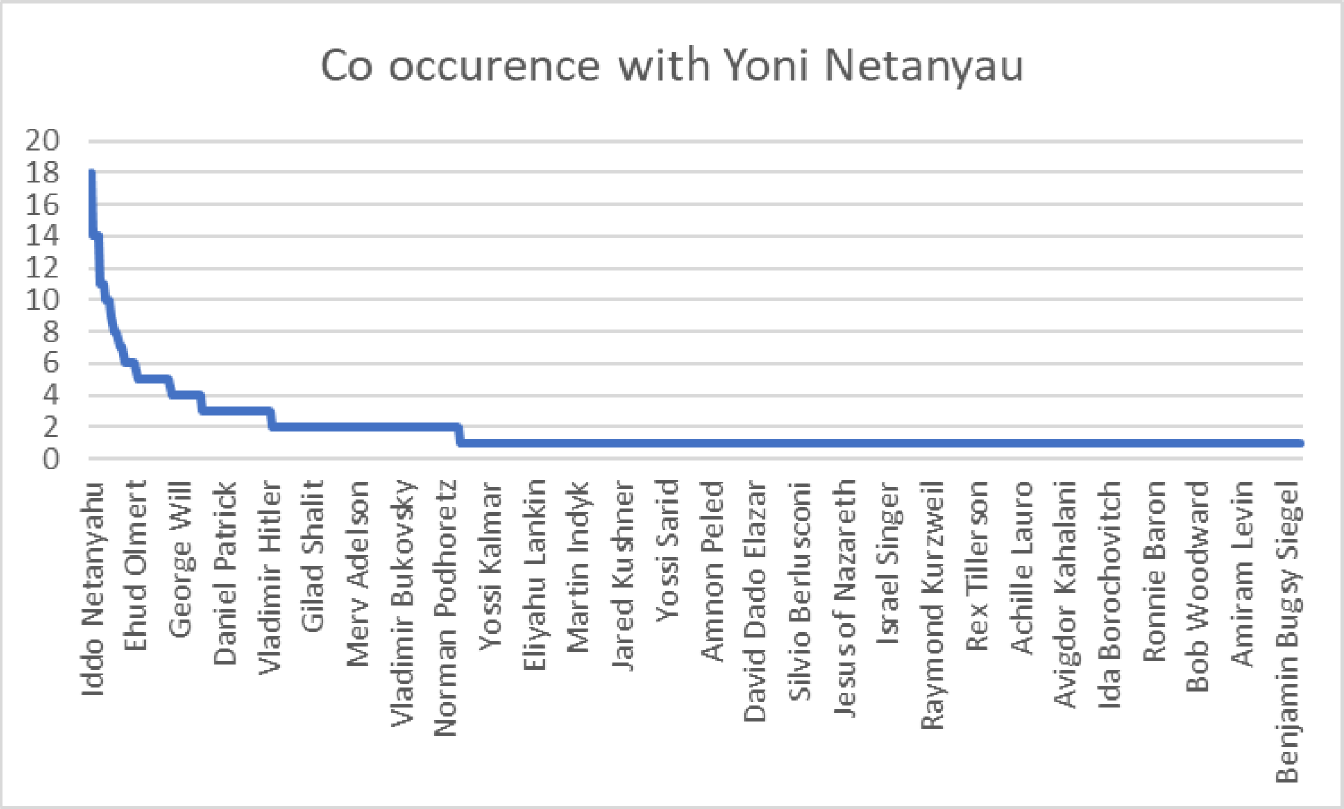
Page rank (based on MENTIONED_IN edges , weighted)

#	name	title	score
1	20	Father	5.46
2	29	Ambassador 1984–1988	4.11
3	28	Diplomat 1982–1984	4.11
4	26	Terrorism 1976–1980	3.65
5	38	First Skirmish 1996	3.60
6	30	Politics 1988–1993	3.05
7	32	Leader of the Opposition	2.68
8	39	Wye River 1998	2.67
9	19	Hasbara 1973–1976	2.45
10	72	New Path to Peace 2020	2.41

*Intensive chapters in terms of persons and events that mentioned in these chapters

Distribution of co occurrence between Yoni Netanyahu to other persons

Name	# of co-occurences
Iddo Netanyahu	18
Benzion Netanyahu	14
SHIMON PERES	14
Yitzhak Rabin	14
Zeev Jabotinsky	11
Sara Netanyahu	11
Menachem Begin	10
Ariel Sharon	10
Ben Gurion.1	9
Ehud Barak	8



Nodes similarity

- The Node Similarity algorithm compares a set of nodes based on the nodes they are connected to. Two nodes are considered similar if they share many of the same neighbors. Node Similarity computes pair-wise similarities based on the Jaccard metric, also known as the Jaccard Similarity Score.
- Given two sets A and B, the Jaccard Similarity is computed using the following formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Persons similarity
(based on CO_OCCUR_WITH edges)

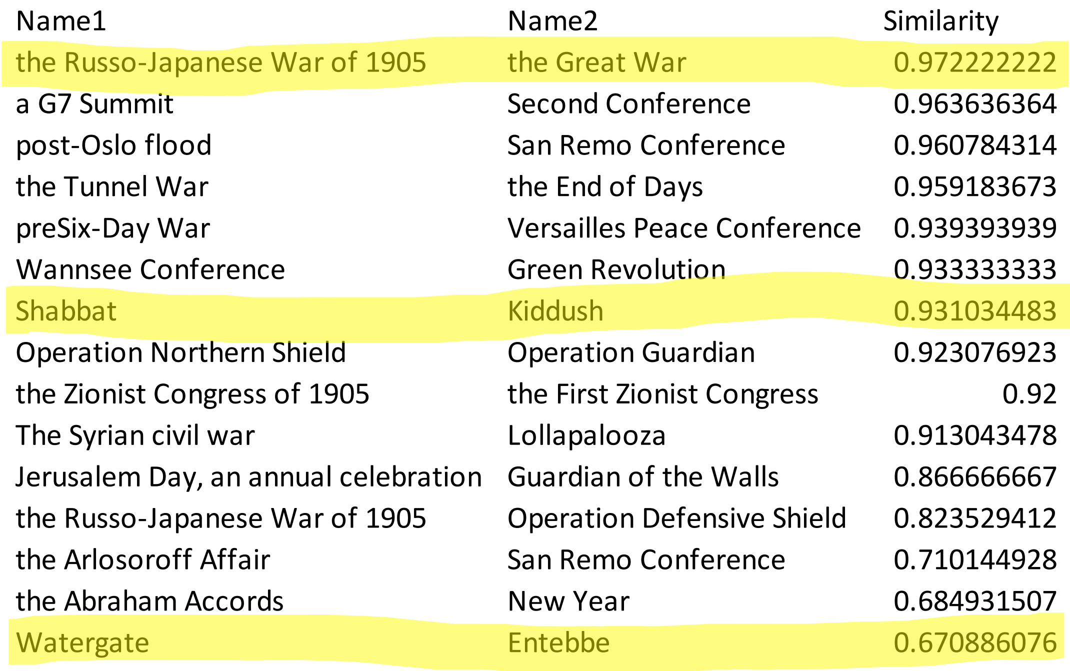
Top 40 similarities

Name1	Name2	Similarity
Paul Nitze	Elmo Bud Zumwalt	97.92%
Peter Lubin	Elie Kedourie	97.75%
Moshe Katsav	David Levy	97.67%
Warren Christopher	Arthur Finkelstein	97.33%
William Ormsby-Gore	Arthur Schlesinger Jr.	97.22%
Jared Kushner	David Friedman	97.06%
Leah Rabin	Aaron Miller	96.77%
William Safire	Bernard Lewis	96.36%
Yelena Bonner	Ahmed Jibril	96.30%
Yossi Sarid	Alexander the Great	96.15%
Yossi Beilin	Alexander Zeid	96.08%
Yitzhak Itzik Molcho	Al Buraq	95.92%
Vladimir Nabokov	Aldo Moro	95.83%
Yaakov Amidror	Boyko Borisov	95.45%
Ronald Lauder	Abu Allah	94.74%
Sleepy Joe.	Abdel Fattah al-Burhan	94.44%
the Thin Man	Ori Yogev	94.12%
alte kaker	Al Gore	93.94%
Warren Buffett	Alex Ferguson	93.75%
Zvi Marom	Amnon Goldenberg	93.55%
Nancy Pelosi	Avshalom Kor	93.33%
the Vilna Gaon	Binyamin Ronn	93.10%
Zalman Shazar	Alex Davidi	92.86%
Tony Blinken	Eyal Yifrah	92.31%
Yossi Cohen	Alfred Dreyfus	92.00%
Zohar Linik	Ahmad Shukeiri	91.30%
William Ormsby-Gore	Abba Eban	90.91%
Ron Kampeas	Hiroo Onoda	90.48%
Zachary Baumel	Aharon Mordechai Freeman	90.00%
Zalman Shazar	Avi Dichter	89.66%
[Pinchas] Bukhris	Amos Goren	89.47%
Yoram Lass	Albert Bourla	88.89%
Yaakov Neeman	Albert Einstein	88.24%
Ramat Shlomo	David Axelrod	87.50%
Shlomo Mordechai	Avichai Mandelblit	86.67%
Yoram Cohen	Angel Gurra	85.71%
Jeane Kirkpatrick	George Schultz	84.68%
Xi Jinping	Adolf Eichmann	84.62%

Events similarity

(based on CO_OCCUR_WITH edges)

Top 15 similarities

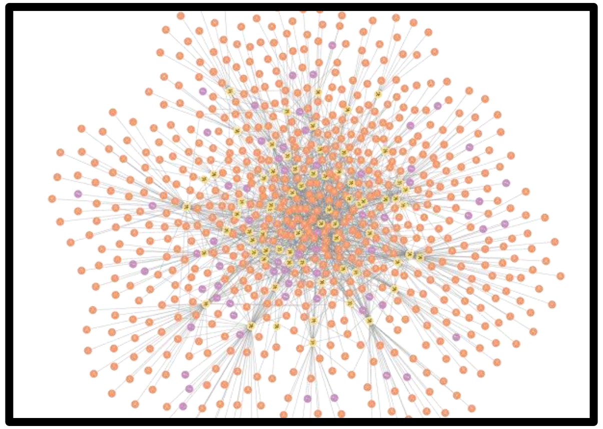


Name1	Name2	Similarity
the Russo-Japanese War of 1905	the Great War	0.972222222
a G7 Summit	Second Conference	0.963636364
post-Oslo flood	San Remo Conference	0.960784314
the Tunnel War	the End of Days	0.959183673
preSix-Day War	Versailles Peace Conference	0.939393939
Wannsee Conference	Green Revolution	0.933333333
Shabbat	Kiddush	0.931034483
Operation Northern Shield	Operation Guardian	0.923076923
the Zionist Congress of 1905	the First Zionist Congress	0.92
The Syrian civil war	Lollapalooza	0.913043478
Jerusalem Day, an annual celebration	Guardian of the Walls	0.866666667
the Russo-Japanese War of 1905	Operation Defensive Shield	0.823529412
the Arlosoroff Affair	San Remo Conference	0.710144928
the Abraham Accords	New Year	0.684931507
Watergate	Entebbe	0.670886076

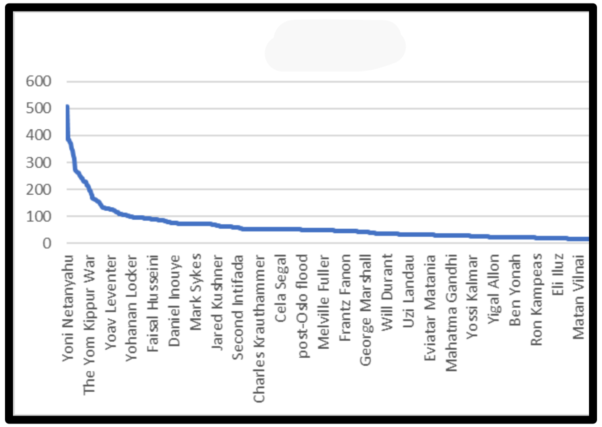
Recap

We **automatically extracted persons and events** via language models.

We automatically performed **clustering with high accuracy**.



We represented the extracted entities as **a social network**



We analyzed the network, to achieve the following insights:

- **Cardinal persons**
- **Cardinal events**
- **Communities**
- **Intensive chapters**
- **Co-occurrence distribution**
- **Persons similarity**
- **Events similarity**

Thanks!