MATINE Final Seminar, November 18, 2015

SemPro: Use of Semantic Mining and Profiling in Building a Situational picture from Social Media and Big Data

(January – November 2015, 50 000 €)

University of Eastern Finland, Department of Environmental Science:
professor Mikko Kolehmainen, research manager Mauno Rönkkö,
researcher Markus Stocker

in cooperation with:
Finnish Defence Research Agency at Riihimäki

MATINE

UNIVERSITY OF EASTERN FINLAND

Contents

1. Research Problem
2. Twitter Data
   - Materials and Methods
   - Results
   - Exploitation
3. Carrot^2 Data
   - Materials and Methods
   - Results
   - Exploitation
4. Future Work

1. Research Problem (1/2)

- Can semantic mining and profiling be used in building a situational picture from Social Media and Big Data?
- Can association networks identify and explain correlations in Social Media and Big Data?
- Topic questions to be targeted:
  - What kind of effect international crises have to Finnish Defence?
  - What external actors and factors affect the discussion about Finnish Defence?
  - What external actors and factors affect the discussion about Finnish-Nato cooperation?

1. Research Problem (2/2)

Hypotheses regarding the research question:

1. *Language used in Social Media and in Big Data has sufficient semantic content.* The language is often colorful, biased, contains acronyms, and incomplete sentences are used.
2. *Reoccurring themes show up as correlations.* Capture them as association networks and visualized as multidimensional graphs.
3. *Association networks help in detecting weak signals.* Requires domain expertise and cannot be fully automated; unexpected correlations can indicate potential weak signals.
4. *Semi-automated methods support co-learning.* Learning based on association networks and visualization.

2. Twitter Data: Materials and Methods

- Twitter data collection
  - March 9 - October 5, 2015
  - Filter for 199 two-word phrases “ukraine russia”
  - Streamed JSON data managed by MongoDB
- Twitter data analysis
  - Descriptive statistics
  - Text mining
  - User profiling and situational picture
  - Three international crises: Ukraine-Russia, Greece-EU, Syria-EU
  - Syria-EU was not filtered
    - Can we detect the Syria-EU signal nonetheless in the data?

2. Twitter Data: Materials and Methods Descriptive statistics

- For each studied collection and geo-located sub-collection
  - Total number of tweets
  - Total number of distinct users
  - User with maximum number of tweets
  - Trend over time for daily count of tweets

2. Twitter Data: Materials and Methods
Text mining

- Most frequent terms associated with each crisis
- For each crisis, topic model for the trend over time
- Performed on random sample
  - Algorithms do not scale to the full collections on 8 GB RAM

2. Twitter Data: Materials and Methods
User profiling

- Clustering users in engaging and non-engaging
  - Structural features of tweets
  - Features such as original tweet length, starts with @, number of @s, number of hashtags, number of links, number of words, normalized tweet length
  - Hypothesis: engaging users more often have tweets that start with @, greater number of @s, lower number of hashtags, lower number of links, greater number of words, and greater normalized text length

2. Twitter Data: Materials and Methods Situational picture

- Map visualization of user profiles
- Integrated view of user profile characteristics
  - Engagement of user
  - Tweet and retweet volumes
- User is geo-located by marker on a world map
- Different visual marker attributes for different characteristics
  - Color, border size, transparency

2. Twitter Data: Materials and Methods
Developed software components

- Collector
  - Persistent query to Twitter API defining the filter
  - Obtains streamed JSON documents
  - Persists JSON documents to MongoDB
  - Operates continuously
  - Implemented in Java
  - Uses Twitter4J library
- Analyser
  - Automated analysis of data collections
  - Implemented in R
  - Using RStudio

2. Twitter Data: Results
Descriptive statistics

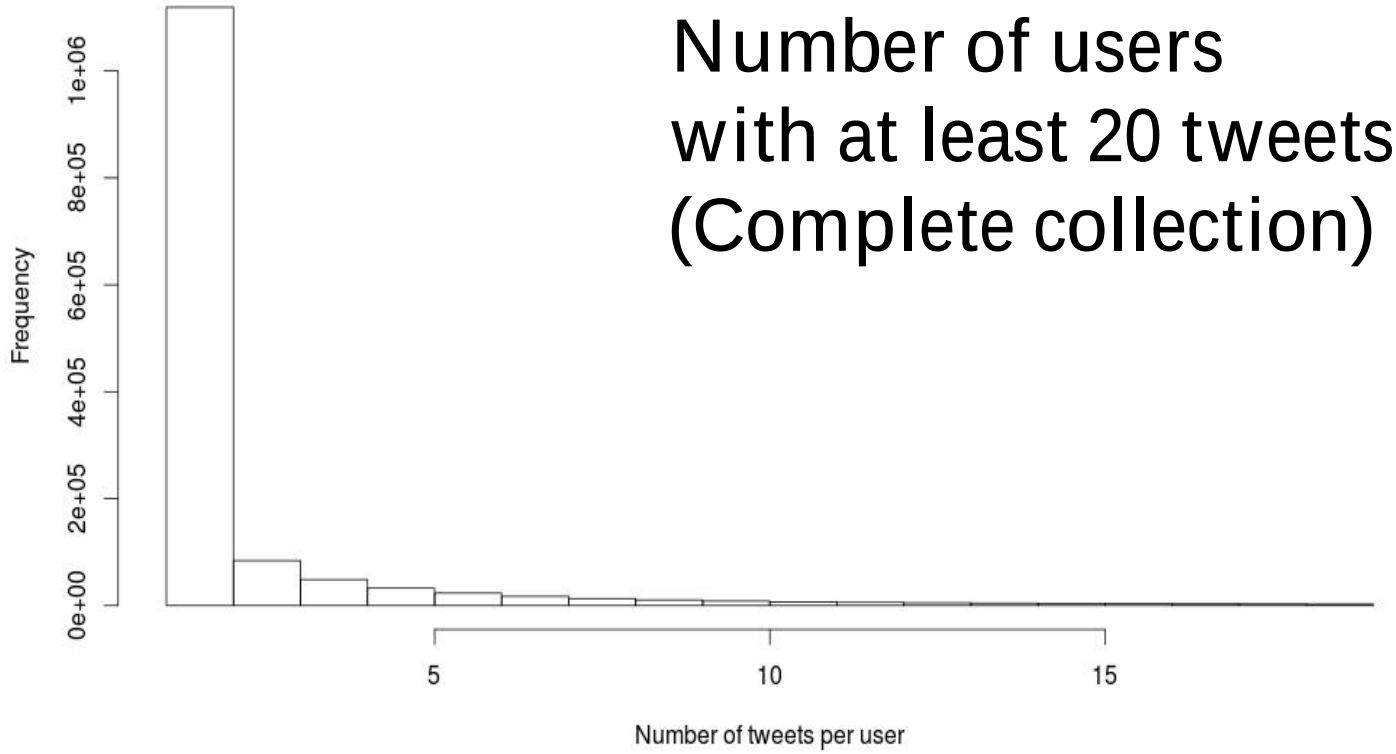
- Complete collection
  - 8 million tweets
  - 1.5 million distinct users
  - Russian news agency (@News18ru) is the most active user
    - The account was suspended by Twitter
  - 0.17% by 6 thousand users are geo-located
  - Most active geo-located user is Paul Erickson (@epaulnet) from CA

2. Twitter Data: Results
Descriptive statistics

	Ukraine-Russia	Greece-EU	Syria-EU
# of tweets	1,285,128	1,445,620	42,892
Distinct users	215,520	351,660	21,015
Max # of tweets	ATO_PRO (48,393)	dlgreece (3,005)	HauteLifestyle (5,839)

Geo-located			
# of tweets (GL)	2,595 (0.20%)	2,112 (0.16%)	26 (0.06%)
Distinct users (GL)	760	1,027	21
Max # of tweets (GL)	epaulnet (562)	MexicoNoAvanza (99)	pidybi (5)

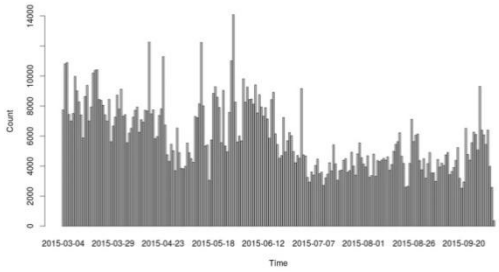
2. Twitter Data: Results
Descriptive statistics



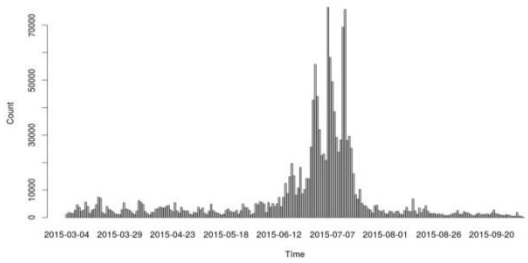
Number of users
with at least 20 tweets
(Complete collection)

2. Twitter Data: Results
Descriptive statistics

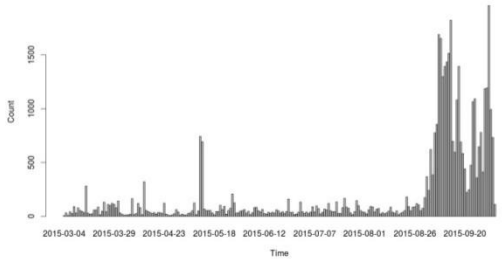
Daily tweets over the entire collection period



Ukraine-Russia



Greece-EU

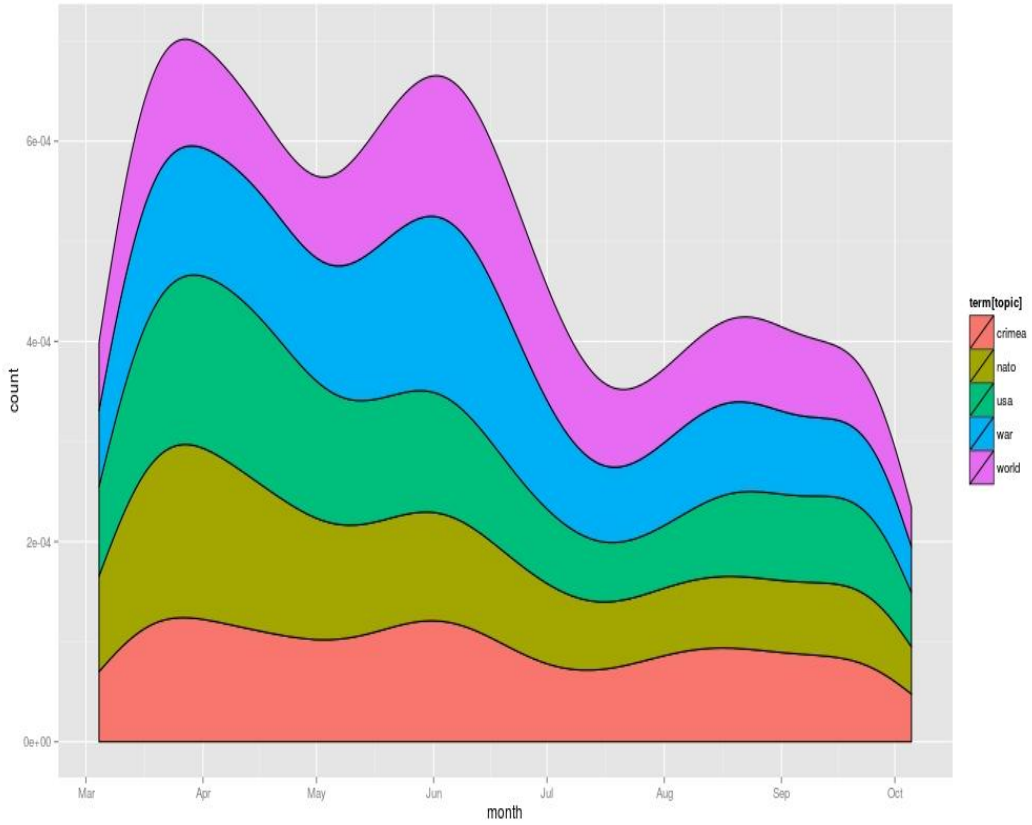


Syria-EU

2. Twitter Data: Results
Text mining

- Most frequent words associated with each crisis
  - Ukraine-Russia: putin, war, nato, world, usa, stoprussianaggress
  - Greece-EU: germani, deal, debt, bank, europ, imf, bailout
  - Syria-EU: isi, russia, ukrain, obama, assad, putin

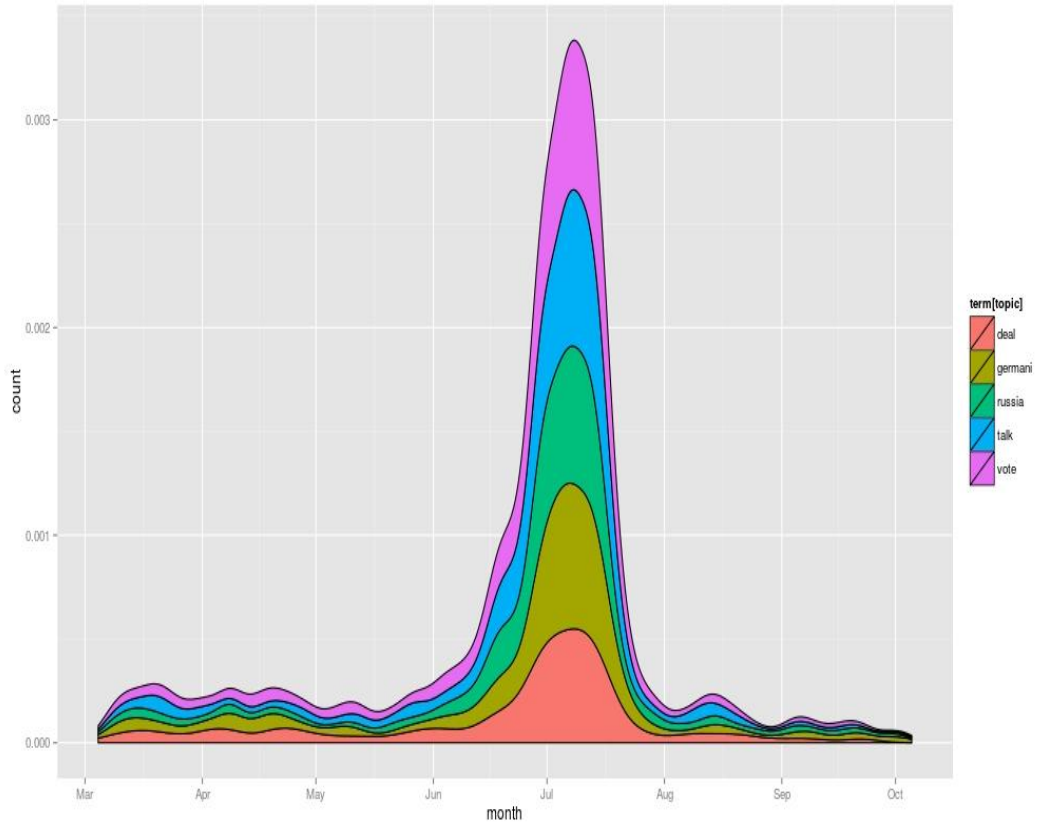
2. Twitter Data: Results
Text mining: Topic model



Russia-Ukraine:

- Crimea
- Nato
- USA
- War
- World

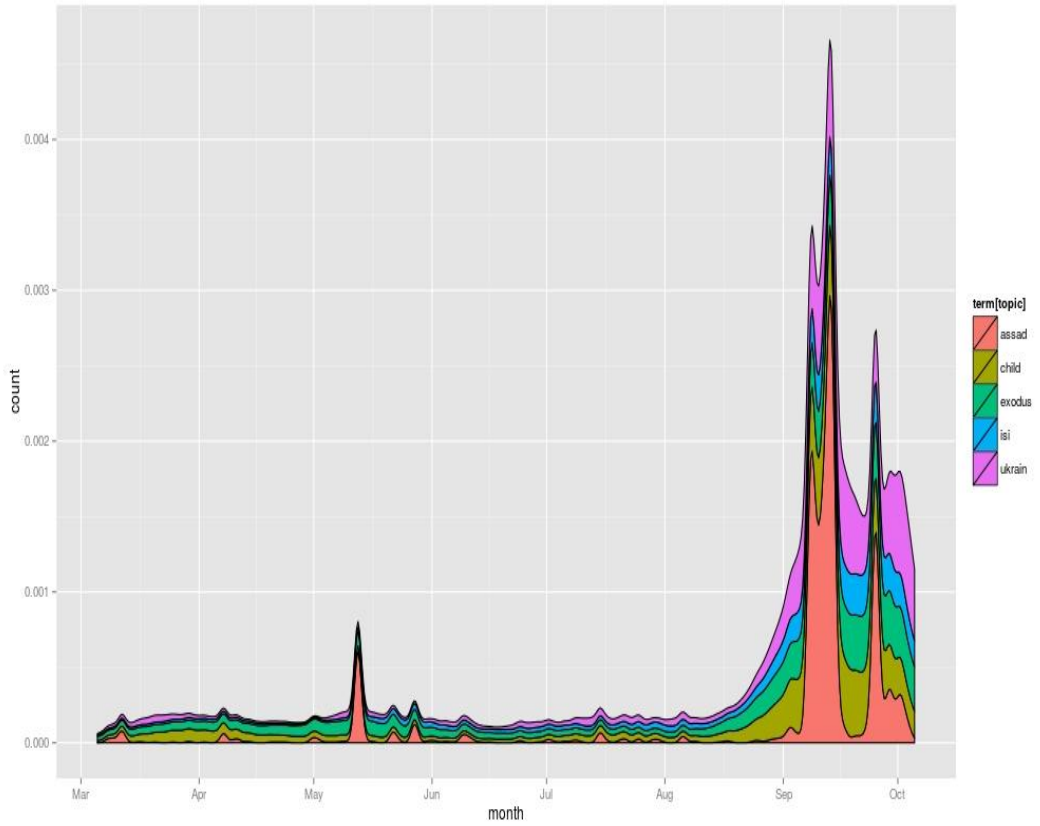
2. Twitter Data: Results
Text mining: Topic model



Greece-EU:

- Deal
- Germany
- Russia
- Talk
- Vote

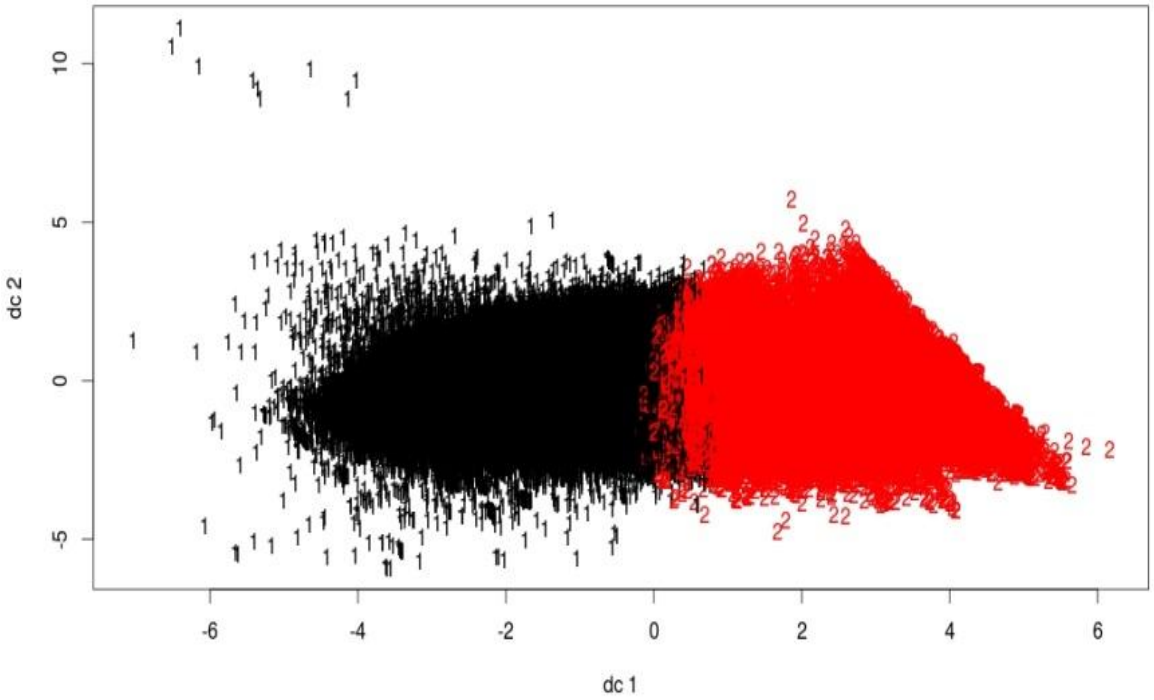
2. Twitter Data: Results
Text mining: Topic model



Syria-EU:

- Assad
- Child
- Exodus
- ISIS
- Ukraine

2. Twitter Data: Results
User profiling: Engagement clusters



2. Twitter Data: Results
User profiling: Cluster centroids

Ukraine-Russia

cluster	length	@...	@s	#	links	words	norm length	Interpretation
1	133	0.7	1.0	1	0.9	8	41	Engaging
2	116	0.3	0.4	4	1.2	3	16	Non-engaging

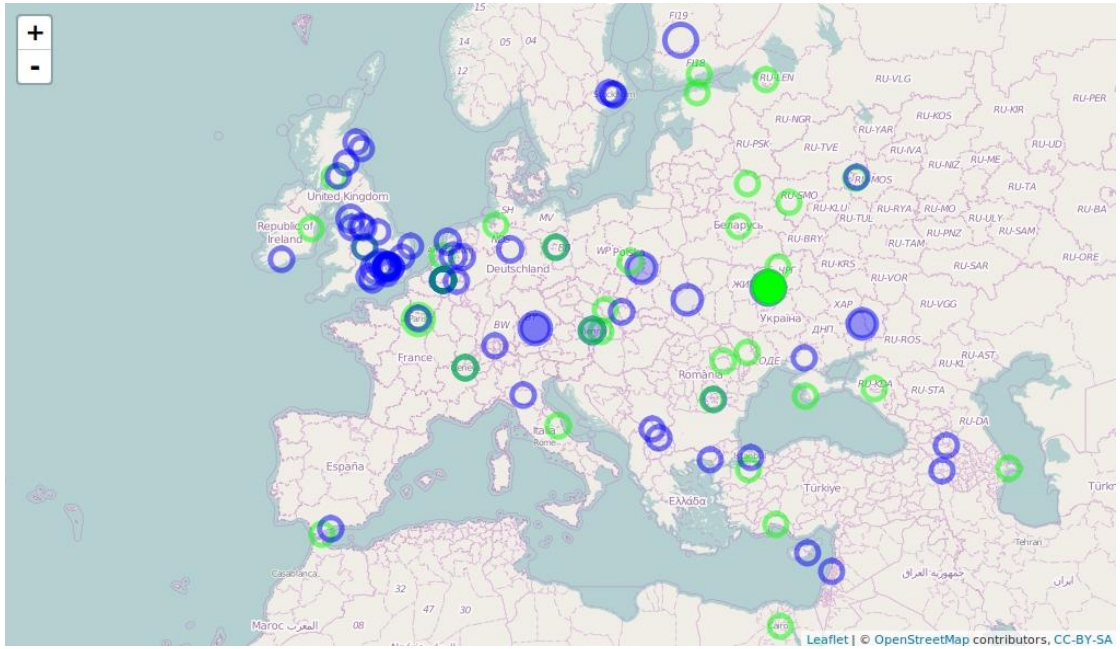
Greece-EU

cluster	length	@...	@s	#	links	words	norm length	Interpretation
1	103	0.3	0.4	1.6	1.6	5	25	Non-engaging
2	133	0.7	0.9	0.9	0.8	10	46	Engaging

Syria-EU

cluster	length	@...	@s	#	links	words	norm length	Interpretation
1	119	0.4	0.6	4.7	1.1	5	25	Non-engaging
2	136	0.5	0.9	0.9	0.8	10	50	Engaging

2. Twitter Data: Results
Situational picture: Ukraine-Russia



Demo: http://enviapps.uef.fi/sempro/ukraine-russia-map.html

2. Twitter Data: Exploitation

- What kind of effect international crises have to Finnish Defence?
  - Use Twitter to discover possible threats from social media
  - Fear a third world war could spark from Ukraine-Russia crisis
  - Localized crisis could involve new actors, including Finland
  - Crisis seen as a power-play (“stoprussianaggression”)
    - Russian muscle flexing affects Finnish Defense
  - Traditional weight of few actors, e.g. USA and NATO
    - Could force countries to take a stance/side/action
  - Economic sanctions (“banrussiafromswift”) effects Finland
    - Also Greek bailouts, decisions by EU/Germany/IMF/ECB
    - Finland tied to decision of more powerful countries
  - Refugee crisis (“exodus”) implications to national security (extremism)

3. Carrot^2 Data: Materials and Methods (1/2)

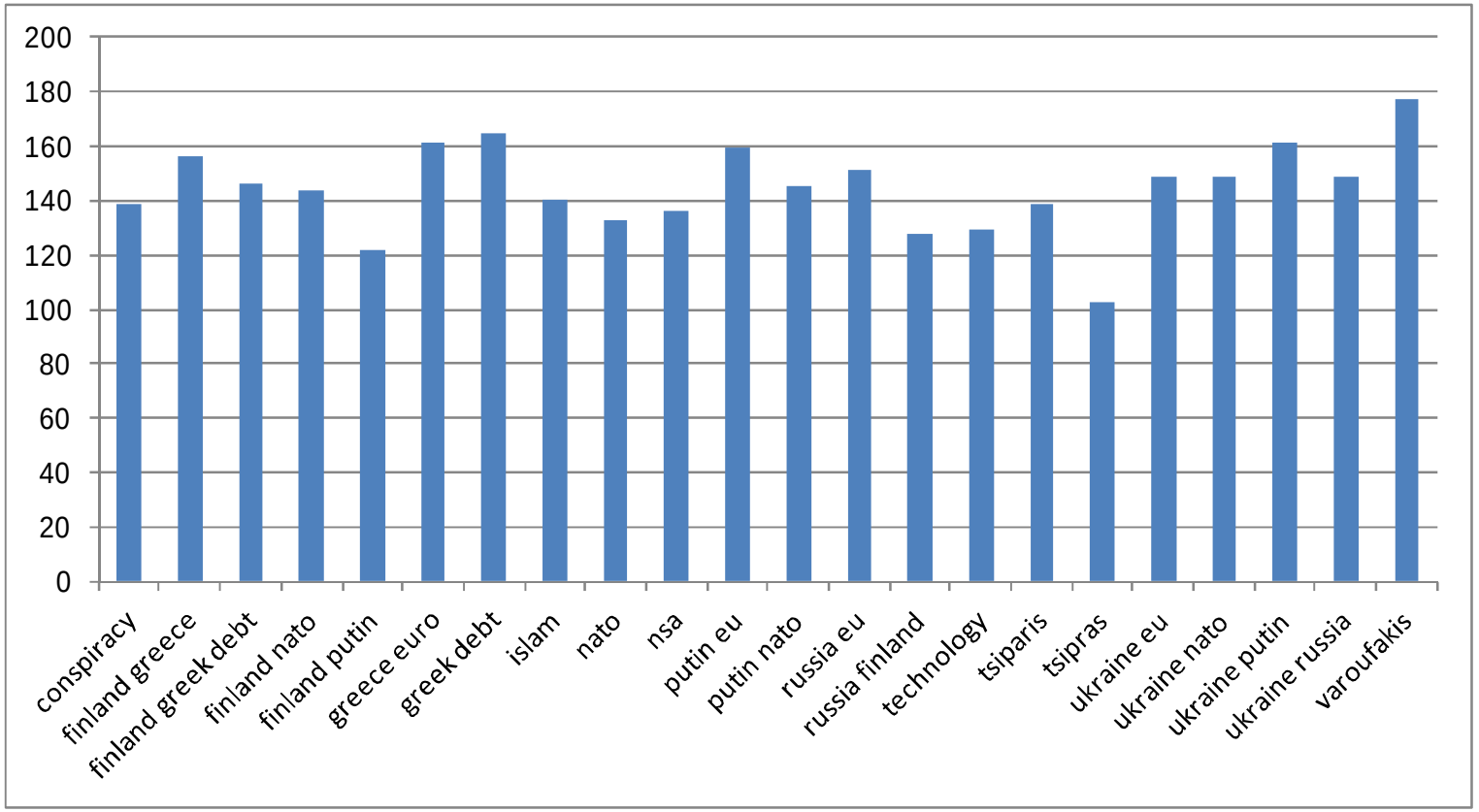
- Carrot^2 service: provides data from Google, Bing, Wikipedia, and News services
- Collected data during April and September 2015
- Search phrases were chosen based on the political state in February - March 2015
- 21 search phrases, including Finland Greek Debt, NATO, NSA, Russia EU, and Ukraine EU
- Collected BBC news feeds during June – September 2015
- Applied Carrot^2 Workbench clustering methods to tweets collected with the same search phrases

3. Carrot^2 Data: Materials and Methods (2/2)

- Carrot^2 service provided a list of associated terms as a foam tree
- Terms were then ranked: 1) the hot terms, 2) intermediate terms, and 3) peripheral terms
- Software was developed in Java, process the data
- The analysis method was based on temporal frequency analysis
- Clustering of terms was done manually to support co-learning.

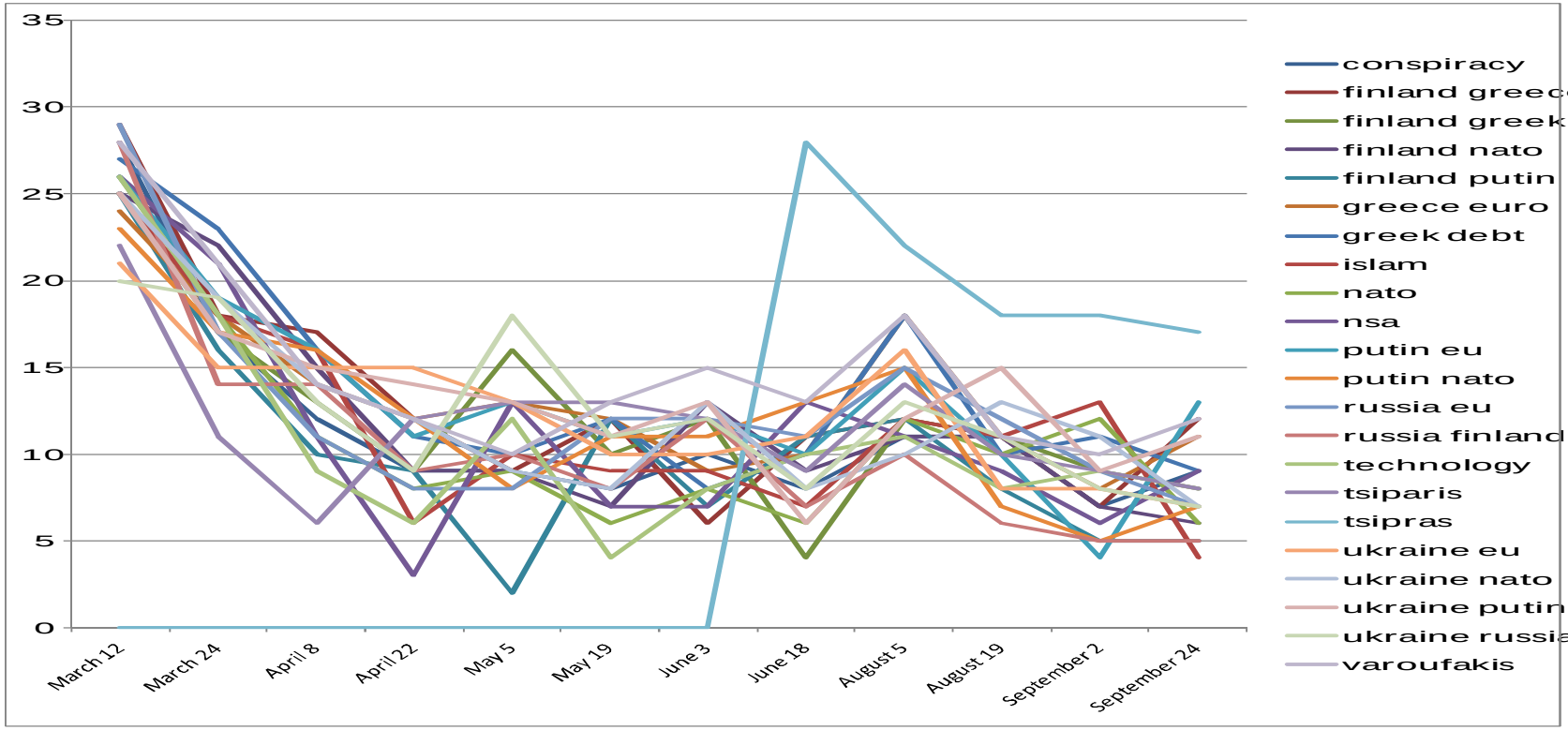
3. Carrot^2 Data: Results

- The total number of unique terms found during April – September 2015 was 3182



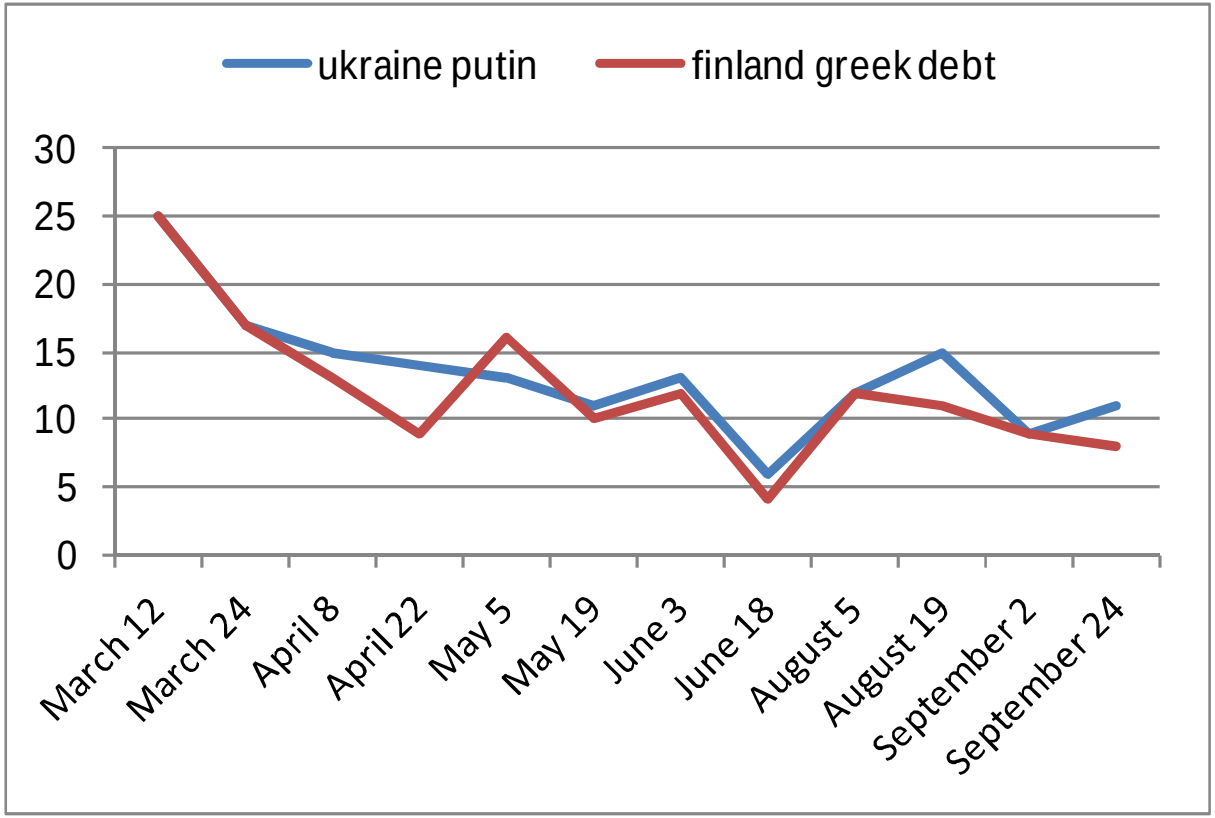
3. Carrot^2 Data: Results

- Number of new terms drops quickly from 25 to about 10 over time
- Because of summer holiday (July), there is a clear increase in the number of new unique terms in the next collection in August



3. Carrot^2 Data: Results

- Unexpected groupping by similarity: “Ukraine Putin” and “Finland Greek Debt”



3. Carrot^2 Data: Results

Observations regarding frequencies of terms:

- There are some interesting and unexpected similarities in the trends
- Trend analysis can be used as a correlation indicator
- The trend analysis can be used to validate correlations
- The trend analysis can be used in search of common factors and causalities

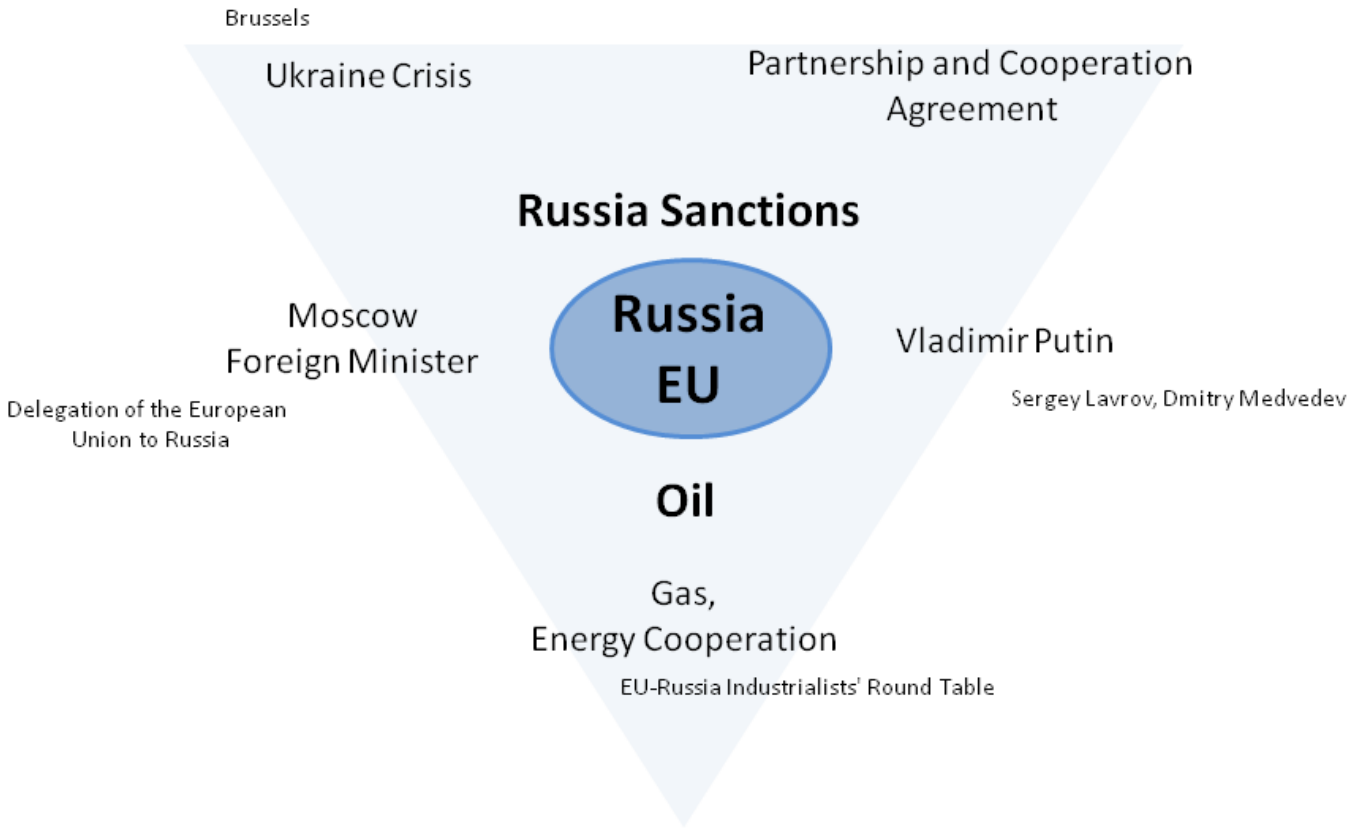
3. Carrot^2 Data: Results

Detailed terms collected for a search phrase example: Russia EU

- Ranking 1: Oil, Sanctions against Russia
- Ranking 2: Crisis in Ukraine, Energy Cooperation, Foreign Minister, Gas, Moscow, Partnership and Cooperation Agreement, Trade war with Russia, Vladimir Putin
- Ranking 3: Annexation of Crimea, Brussels, Dairy Products, Delegation of the European Union to Russia, Dmitry Medvedev, EU-Russia Industrialists' Round Table, Europe's Russia Denial, European Atomic Energy Community and the European, Interested in Modern Russia and the Future, Limits and Necessity of Europe's Russia Sanctions, Russia's European Delusion, Russia's Exit from Europe, Russia's Gas Pipeline Strategy and Europe's Alternatives, Russia's Travel Ban, Sergey Lavrov

3. Carrot^2 Data: Results

- Association network example before June: Russia EU

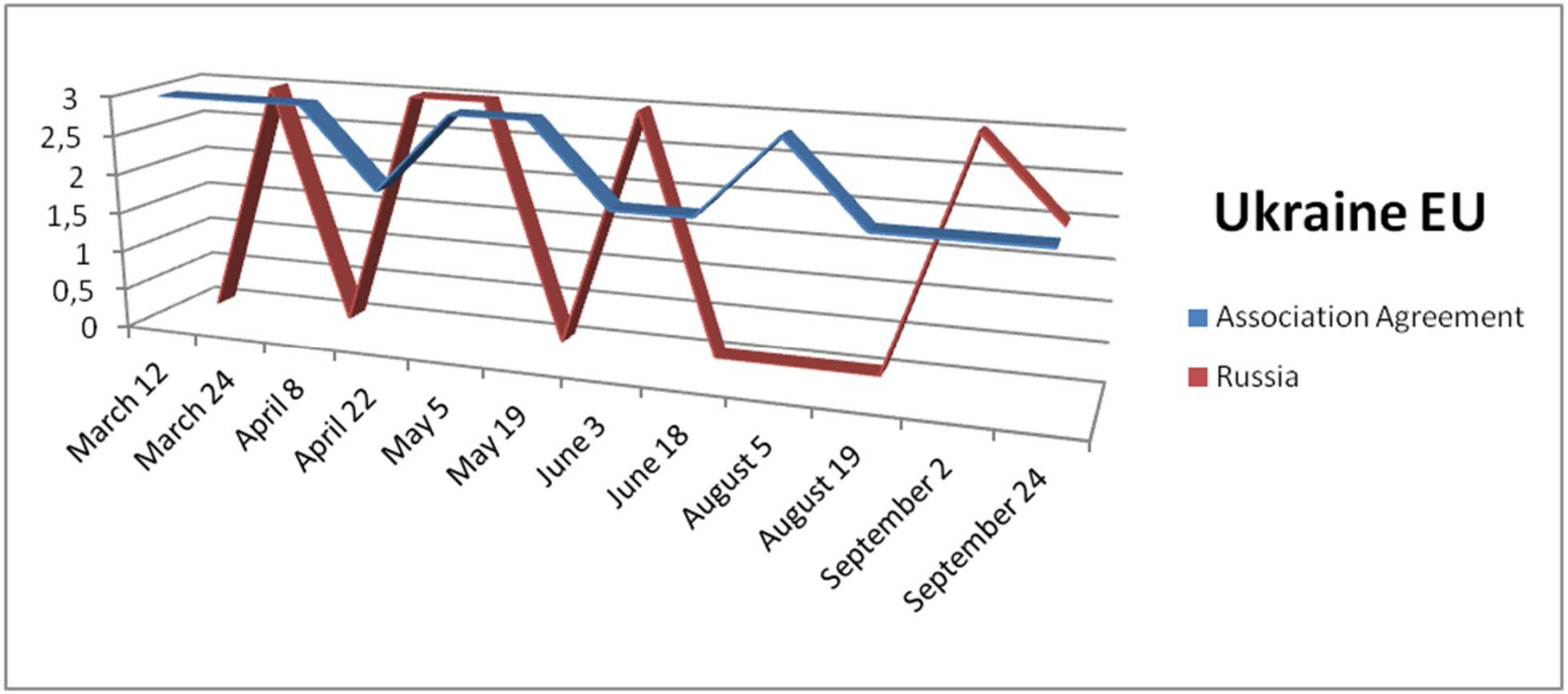


3. Carrot^2 Data: Results

- Association network development example after June: Russia EU

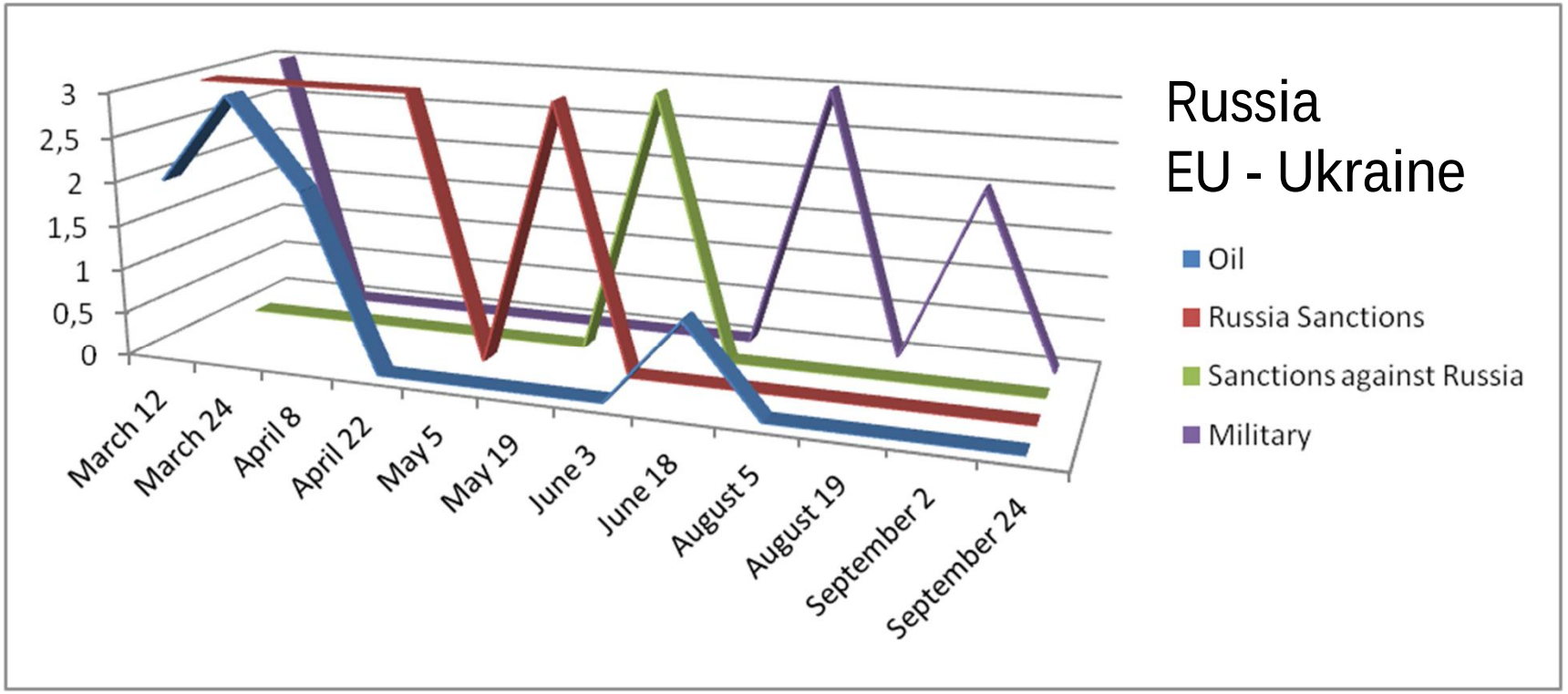
3. Carrot^2 Data: Results

- Example of term ranking trends:



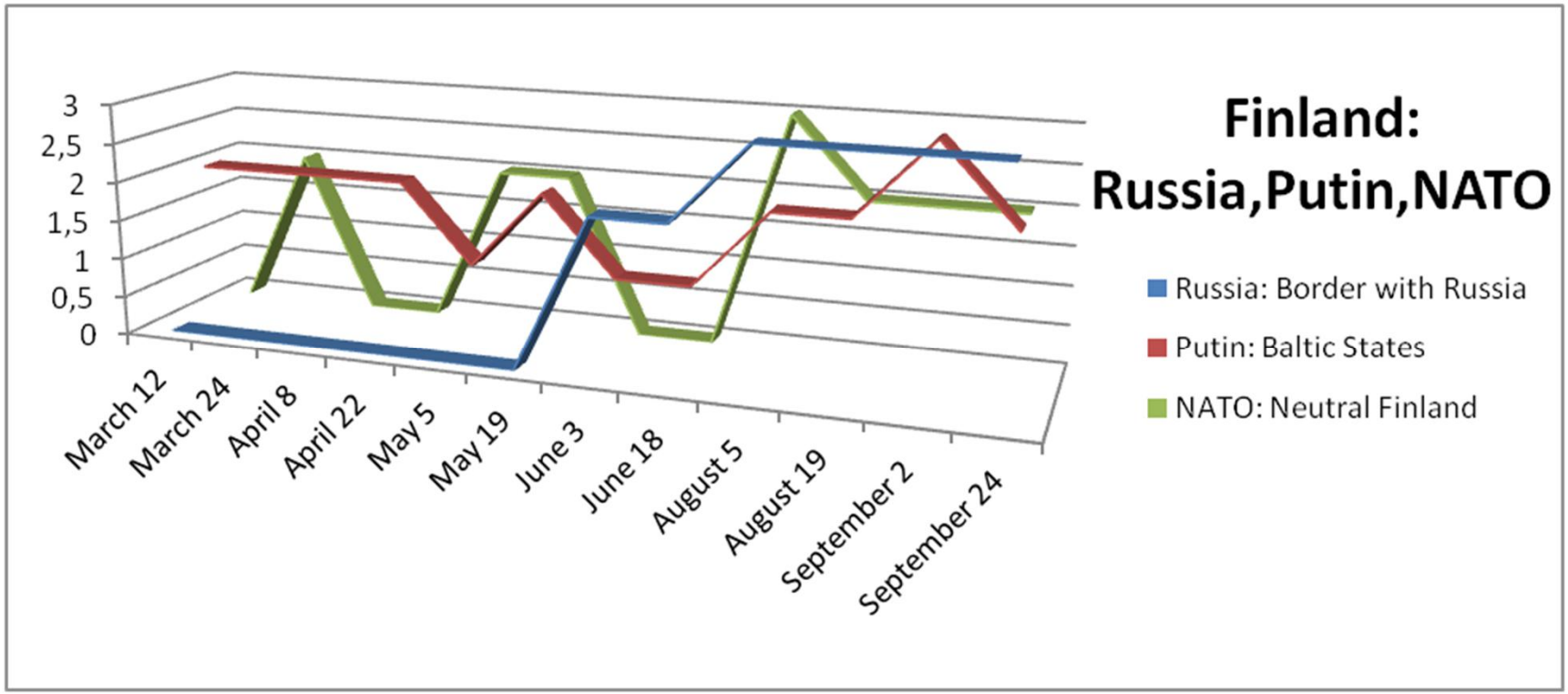
3. Carrot^2 Data: Results

- Example of term ranking trends:



3. Carrot^2 Data: Results

- Example of term ranking trends:



3. Carrot^2 Data: Results

BBC News Headlines correlations:

1. Finland does not really exist; emphasis by a news agency!
2. “Oil”, “Gas”, and “Energy Cooperation” are not current topics for BBC; significance for UK?
3. Long term events are covered by BBC News, such as “Ukraine Crisis”
4. Terms like, “Sanctions against Russia”, “Annexation of Crimea”, and “Russia's Travel Ban” co-appeared through both means of collection in a timely fashion; reflecting internationally significant events?
5. Terms like “Migration”, ”Georgia”, “Syria”, and “Google” did not appear through Carrot^2 at all; topics reasoned and put into a context by a journalist?

Conclusion: The news headline can also be used to validate that some term found through Big Data is related to an event of importance.

3. Carrot^2 Data: Results

Analysis of tweets by using Carrot^2 Workbench:

- Applied the workbench and its clustering methods to subsets of tweets collected with a specific phrase
- We clustered five subsets of tweets containing the words “Russia” and “EU”
- The subsets contained 100, 500, 1000, 2000 and 4000 tweets
- The tweets were further merged based on the sender
- Removed all tweets that were similar to one another

3. Carrot^2 Data: Results

Example: 100 tweets clustered on search phrase “Russia EU”:

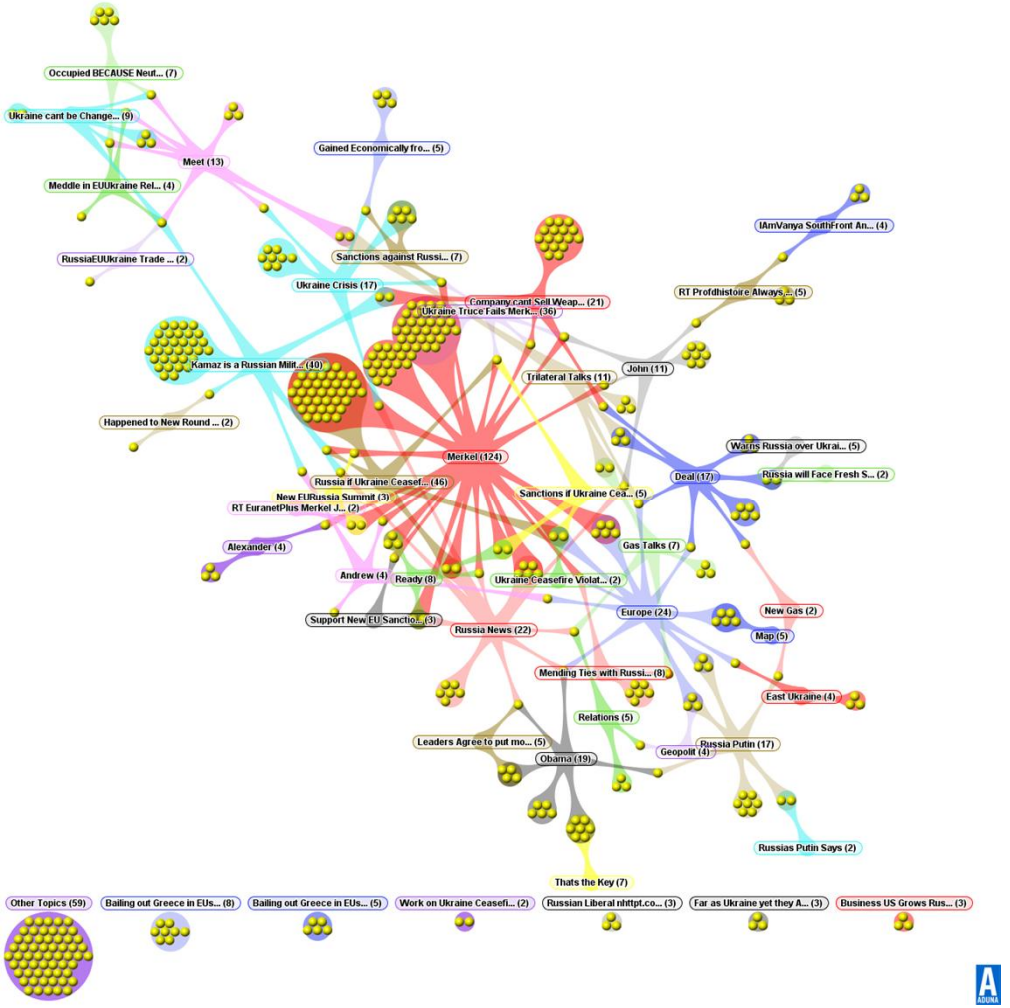
3. Carrot^2 Data: Results

Example: 1000 tweets clustered on search phrase “Russia EU”:

3. Carrot^2 Data: Results

Interdependencies between clustered tweets can then be visualized as a cluster map.

However, this works only for a limited number of tweets. For many thousands of tweets the map becomes unreadable.



3. Carrot^2 Data: Results

Conclusion about the analysis of tweets by using Carrot^2 Workbench:

- The Workbench should be used on small subsets of tweets
- The terms found through tweets are more explanatory in comparison to those found from Big Data with the Carrot2 service
- Does not work equally well for all kinds of search phrases
- The content and quality of tweets is of significance when using tools such as Carrot^2 Workbench

3. Carrot^2 Data: Exploitation (1/4)

Topic questions to be answered:

- What kind of effect international crises have to Finnish Defence?
- What external actors and factors affect the discussion about Finnish Defence?
- What external actors and factors affect the discussion about Finnish-Nato cooperation?

3. Carrot^2 Data: Exploitation (2/4)

What kind of effect international crises have to Finnish Defence?

1. The borderline of Finland becomes a focal topic.
2. Effects on weapons acquisition.
3. Relations to alliances and organizations.
4. Participation in support actions.
5. Effect of trade wars and sanctions.
6. Energy availability and self-sufficiency (Energy security).
7. Public relations and international debate / Information warfare

3. Carrot^2 Data: Exploitation (3/4)

What external actors and factors affect the discussion about Finnish Defence?

1. International affairs (Finland, EU, Russia, NATO)
2. EU actions. (sanctions, travel bans, and trade wars)
3. Lack of shared internationally interests. (BBC)
4. Individuals. (president Putin and Chancellor Merkel)
5. Propaganda.
6. Elections.
7. Geographics. (border and arctic dimensions)
8. Energy security.
9. Refugee crisis.

3. Carrot^2 Data: Exploitation (4/4)

What external actors and factors affect the discussion about Finnish-Nato cooperation?

1. Russia. (relationships in Europe, borderline of Finland)
2. Ukraine crisis.
3. EU relationships.
4. Baltic States.
5. Image of neutral Finland.
6. Sweden. (“Sweden and Finland's Awkward NATO Tango”)
7. Politicians and elections.
8. Military exercises.

4. Future Work

- Twitter Data
  - Big data analysis (no sampling) with Hadoop, SPARK, etc.
  - Streamed data analysis for real-time situational picture
- Carrot^2 Data
  - Development of Bayesian networks based on association networks
  - Analysis of country dependent bias in the association networks

Contact information:
Mauno Rönkkö
mauno.ronkko@uef.fi
puh. +358 40 355 2202

UNIVERSITY OF EASTERN FINLAND

MATINE

www.uef.fi