# Identifying Discordant Cybersecurity Terms Between English and Russian Speakers
CS224W: Final Paper

Ashe Magalhaes

***Abstract*— This paper aims to mitigate the consequences of false attribution and ensuing crisis escalation between the U.S. and Russia by examining a tool that identifies the most discordant cybersecurity-related words used between English and Russian speakers. We leverage the GDELT database to create co-occurrence networks of cybersecurity terms, conduct additional graph processing, and compute network distance scores for the English and Russian translation of each term.**

**The result is a list of the most frequently used cybersecurity terms across all content published in the last year in order of discordance. We further analyze our results by considering community structure and the features that drive network distance. The contribution of the paper is a novel approach to U.S.-Russian cooperation in cyberspace that establishes a foundation for understanding the differences in how English and Russian speakers contextualize cybersecurity.**

## I. INTRODUCTION

According to a survey by the Pew Research Center released in October 2014, about two-thirds of the technology experts polled expected a major cyberattack in the world by 2025, which is likely to lead to significant loss of life or property losses of tens of billions of dollars [1]. Because offensive cyber operations necessarily place a heavy and asymmetric burden on a passive defense posture, the international debate over cybersecurity includes a consideration of attack options for defensive purposes [2]. This, coupled with the difficulty in establishing attribution of cyberattacks, presents the problem of proportional responses directed at a non-adversarial group or nation-state. The U.S. and Russia are particularly vulnerable to conflict escalation resulting from the false attribution of cyber attacks because of heightened tensions and the breakdown of communication channels in October 2016 [3].

The problem of mitigating the threat of false attribution is particularly difficult because ascertaining the machines associated with a malicious cyber incident involves technical forensics, which may take months or longer if identification is possible at all. Even if machine attribution is established, this does not necessarily lead to identification of the state or non-state actor that coordinated the event [4].

Previous work in this area provides solutions for cooperation in cyberspace that involve technical forensics [5] or the development of a legal framework [6]. However, these approaches have distinctly Western applications. In order to promote cooperation between the U.S. and Russia, a solution that considers both American and Russian perspectives must be developed. This solution should include an empirical analysis on publicly available data, so as to prevent bias and develop trust between the two governments.

In this paper, we address the problem of attribution of offensive cyber operations as it relates to the U.S. and Russia by adapting existing algorithms for network comparison to create a tool that identifies discordant cybersecurity terms used by English and Russian speakers. We leverage a Global Database on Events, Language, and Tone (GDELT) to create pairs of co-occurrence networks of cybersecurity terms and related themes from content produced in the U.S. and in Russia. The network pair with the largest network distance metric can be understood as the most discordant cybersecurity term. After identifying discordant cybersecurity term(s), we explore the implications of the results for policymakers. Our goal is to promote cooperation and communication and avoid hostility and communication withdrawal between the U.S. and Russia.

## II. BACKGROUND

This section outlines the GDELT database, the concept of a co-occurrence network, and network distance.

### A. GDELT Database

The GDELT database originated from a desire to better understand global human society and the connection between communicative discourse and large-scale behavior. Its goal is to codify the entire planet into a computable format using all available open information sources [7].

To date, the GDELT database has been used in projects that visualize the past 24 hours of global conflict and protest, perform rapid triaging and assessment of the most important "influencers" in an industry, topic, organization, or geographic region, and provide visibility into global trends and emerging social, political and economic risks.

This research leverages the GDELT 2.0 Global Knowledge Graph (GKG). GKG organizes a catalog of global events in over 300 categories into an annotated metadata graph over the worlds news each day. Totaling over 200 million records and growing at a rate of half a million to a million articles a day, it is perhaps the largest open data graph detailing global human society. It includes the GDELT Translingual feature to provide translation coverage of all monitored content in 65 languages.

### B. Co-occurrence Network

A co-occurrence network allows us to anaylze the connections among terms in a database. For a selected term, the network shows all the terms it is most connected to during a particular time period or in relation to a particular topic. The weight of a given edge in the network is computed as

the count of that term divided by the sum of total counts of all terms. The existing literature on co-occurrence networks includes

- Choosing the word most typical in context using a lexical co-occurrence network [8],
- Using co-occurrence network structure to extract synonymous gene and protein names [9],
- Exploring co-occurrence network patterns in soil microbial communities [10].

### C. Network Distance

Given a set of networks of possibly different sizes, how can we efficiently provide a measure of structural similarity? We build on the work of Berlingerio et. al [11] to arrive at an effective and scalable solution for computing network distance. Berlingerio et. al describe the concept of network similarity (or distance) and how it can be approached by forming graph signatures through feature extraction. After signatures are generated, a distance function can be applied to provide meaningful network distance metrics. Their algorithm is attractive because it gives similarity scores that are size-invariant and has runtime complexity that is linear on the number of edges in the graph. We adapt their *NETSIMILE* algorithm to our implementation by using four of the features they describe, using the aggregators they describe, using a signature vector for each graph, and using the Canberra Distance function. We adapt our signature vector to our networks by considering center node features. We build on the work of Berlingerio et. al by analyzing how features drive network distance.

## III. METHODOLOGY

This section provides an overview of the data collection process, the formulation of co-occurrence networks, additional graph processing, and the design choices for our network distance function.

### A. Data Collection Process

The GDELT 2.0 Global Knowledge Graph (GKG) is available in BigQuery, Google's enterprise data warehouse for analytics. GKG's complex multi-delimiter fields are accessible with SQL but require advanced string manipulation features of BigQuery to generate final data results that are downloadable as a spreadsheet. BigQuery provides users with 1 free TeraByte (TB) every month to query data. Because GKG has over 2.5 TB of events generated from February 2015 to February 2016 alone, we quickly exceeded our quota during the data collection process. For this reason, only two networks are complete graphs in the data processing step of our experiment.

We choose our terms of interest by running a query that returns the top ten themes associated with the term CYBER_ATTACK in the last year. We choose this term after contacting GDELT's creator Kalev H. Leetaru, who recommended we use CYBER_ATTACK and provided documentation that lists all themes. GKG is still in development and there is no documentation that provides more information on a given theme (i.e. it is not completely clear what a theme titled TAX_FNCACT_WRITER means). Based on Kalev's feedback, we operate under the assumption that GDELT will provide this documentation by Spring 2017.

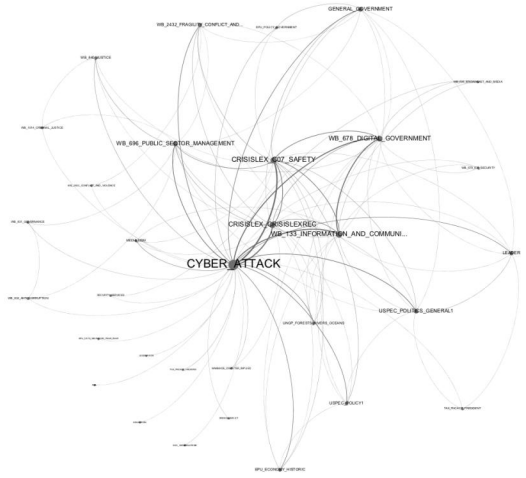



Fig. 1. Co-occurrence Network Visualization of English Term CYBER_ATTACK

In our query, we specify that the count variable should indicate the number of documents that mention a given theme, rather than the number of mentions of a theme (i.e. a theme that appears 100 times in a given article is only counted as one appearance)[12]. This query provides a list of the top 10 cybersecurity-related terms across all content published in the last year, written in over 65 different languages. See Appendix I for exact SQL code.

### B. Co-occurrence Network Formulation

Given this list of terms, we next form a query that returns a co-occurrence network. We do this because it is the connections among themes that is of interest rather than just their frequency of occurrence. This query compiles the list of themes from each matching article and records how many articles each pair of names appear in together. It only considers content published from 11/01/2015 to 11/01/2016 and limits the number of edges to 100 in order to maintain a computationally tractable problem. The final results of this query is a histogram of all pairs of themes and how many articles they have appeared in together—essentially, the edge list of the co-occurrence network of themes [12]. We conduct this query twice, once for content originally written in English and once for content originally written in Russian. For specifying language, we rely on the GDELT Database's Translingual feature which allows us to constrain queries to news coverage in a specific language. The result is a pair of co-occurrence networks from English content and Russian content for each cybersecurity term. See Appendix II for the exact SQL code. See Figure 1 for visualization. Edge thickness is proportional to edge weight and node size is proportional to node degree.

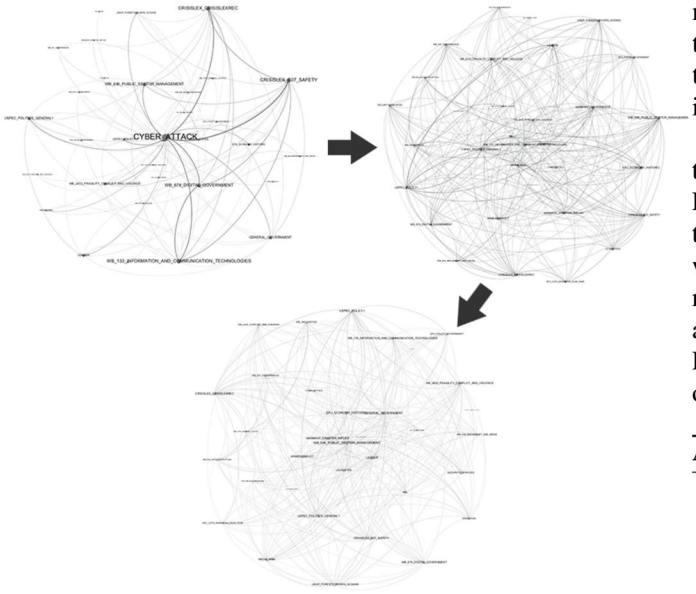



Fig. 2. Graph Processing Visualization of English Term CYBER_ATTACK

After receiving the results of each query, we download the data into a CSV file. We use the python interface for Scalable Network Application Package (SNAP) to load the data into a graph data structure. SNAP is a general purpose, high performance system for analysis and manipulation of large networks.

### C. Additional Graph Processing

We consider that each network has similar structure because they are all co-occurrence networks (i.e. many edges connect a central node to every other node). In order to create distinct networks, we conduct additional graph processing. We begin by focusing on the English and Russian network pair of the term CYBER_ATTACK, the term we used for our initial query to find other cybersecurity-related terms. We make each of those networks a complete graph, or a graph in which every node shares an edge with every other node. To do this, we develop a script that automates the process of iterating through each pair of nodes in the network and querying for their co-occurrence count. We then create an edge between the pair of nodes.

Next, for every other English co-occurrence network, we add edges between nodes if those nodes are linked by an edge in any other English network. For example, if we are looking at the English co-occurrence network of the term GENERAL_GOVERNMENT and find that nodes ARMED_CONFLICT and ELECTION exist but are not linked by an edge, we can search all of our other English networks. If we find the edge and its count associated with ARMED_CONFLICT and ELECTION in another network, we then add that edge to the GENERAL_GOVERNMENT English co-occurrence network. We do the same for the Russian networks.

This technique successfully increases the number of edges in our co-occurrence networks. For our English networks, the number of edges increases to **51%** of the edges present in the complete graph, on average. For our Russian networks, the number of edges increases to **58%** of the edges present in the complete graph, on average.

To further diversify our networks, we find the average of the edge counts for each network and use that as a threshold. For each edge in a network, if the edge count is less than our threshold, we delete it from the graph. The result is a network with edges that represent a significant correlation between nodes, or terms in our dataset. We conduct our analysis on a set of 10 structurally unique network pairs of English and Russian cybersecurity terms. See Figure 2 for a visualization of all data processing.

**Algorithm 1** Generate Signature Vectors

**procedure**
$\{F_{G_1}, F_{G_2}, ..., F_{G_k}\}$ = *GetFeatures*
**for all** $x \in \{F_{G_1}, F_{G_2}, ..., F_{G_k}\}$
    *centerNode = getCenterNode(x)*
    $s_{G_x} = []$
    **for all** *feat*$\in F_{G_x}$
        $s_{G_x} \cup$ {*median(feat), mean(feat), stdev(feat), skewness(feat), kurtosis(feat)*}
    **end for**
    $s_{G_x} \cup$ *getNeighborNo(centerNode)*
    $s_{G_x} \cup$ *getNodeCC(centerNode)*
    $s_{G_x} \cup$ *getAvgTwoHop(centerNode)*
    $s_{G_x} \cup$ *getAvgCC(centerNode)*
**end for**
**return** $\{s_{G_1}, s_{G_2}, ..., s_{G_k}\}$

### D. Network Distance Function

Our network distance function consists of three parts: feature extraction, feature aggregation, and network comparison.

**Feature extraction.** We generate a set of structure features for each node based on its local features [11]. Specifically, we compute the following four features:

- $d_i = |N(i)|$: Number of neighbors (i.e. degree) of node $i$; $N(i)$ denotes the neighbors of node $i$.
- $c_i$: clustering coefficient of node $i$, defined as the number of triangles connected to node $i$ over the number of connected triangles centered on node $i$
- $\bar{d}_{N(i)}$: average number of node $i$'s two-hop away neighbors, calculated as $\frac{1}{d_i}\sum_{j \in N(i)} d_j$
- $\bar{c}_{N(i)}$: average clustering coefficient of $N(i)$, calculated as $\frac{1}{d_i}\sum_{j \in N(i)} c_j$

**Feature aggregation.** After the feature extraction step, we have a set of feature matrices $\{F_{G_1}, F_{G_2}, ..., F_{G_k}\}$. We now generate a single "signature" vector for each graph to produce efficient and effective comparisons. We use the following five aggregators on each feature: *median, mean, standard deviation, skewness, and kurtosis.* For each node, we append these aggregators to a single list, our signature vector.

**Center node features.** After adding the aggregators for each feature, we compute the same features for the center

TABLE I
CYBERSECURITY TERMS

| TERM | COUNT |
|---|---|
| CYBER_ATTACK | 1888964 |
| CRISISLEX_C07_SAFETY | 1251265 |
| CRISISLEX_CRISISLEXREC | 1070679 |
| WB_133_INFORMATION_AND_COMMUNICATION_TECH | 1022871 |
| WB_678_DIGITAL_GOVERNMENT | 991150 |
| WB_696_PUBLIC_SECTOR_MANAGEMENT | 776948 |
| USPEC_POLITICS_GENERAL1 | 733939 |
| GENERAL_GOVERNMENT | 689712 |
| WB_2432_FRAGILITY_CONFLICT_AND_VIOLENCE | 687723 |
| USPEC_POLICY1 | 684035 |

node specifically and append them to the signature vector. Because our networks started off as co-occurrence networks, the center node has a significance that should be represented in our network distance function.

**Network Comparison.** After the aggregation step, we have a signature vector $\vec{s}_{G_j}$ for each graph $G_j \in \{G_1, G_2, ..., G_k\}$. We conduct a pairwise comparison between the English and Russian signature vectors $\vec{s}_{G_{j,ENG}}, \vec{s}_{G_{j,RUS}}$ for each graph. We use the Canberra Distance function,

$$d_{Can}(\vec{s}_{G_{j,ENG}}, \vec{s}_{G_{j,RUS}}) = \sum_{i=1}^{d} \frac{|(\vec{s}_{G_{j,ENG}})_i - (\vec{s}_{G_{j,RUS}})_i|}{\vec{s}_{G_{j,ENG}})_i + (\vec{s}_{G_{j,RUS}})_i}$$

because it is discriminative, a good property for distance measure. This is because the Canberra Distance function is sensitive to small changes near zero, as it normalizes the absolute difference between graphs.

**Computational Complexity** As Belangario et al. [11] demonstrates, the runtime complexity of this algorithm is

$$O(\sum_{j=1}^{k}(fn_j + fn_j log(n_j)))$$

where $k$ is the number of graphs, $n_j$ the number of nodes, and $f$ the number of structural features extracted. In real world networks,

$$f << n_j << m_j$$
$$n_j log(n_j) \approx m_j$$

where $m_j$ is the number of edges. In other words, the algorithm is linear on the number of edges when used on our GDELT database.

## IV. RESULTS AND ANALYSIS

This section provides the empirical results and evaluations of our experiments.

**Cybersecurity Terms.** We began by finding the top ten cybersecurity terms across all languages. They are enumerated in Table 1 along with their counts. Terms that include information and communication, digital government, public sector management, general politics, general government, conflict and violence, and policy are expected, as cybersecurity policy has become a major area of concern for both governments in the past year. The terms that are less clear in their meaning are CRISISLEX_C07_SAFETY and CRISISLEX_CRISISLEXREC. We hypothesize that these terms originate from `CrisisLex.org`, a tool that recommends crisis lexicon for social media analysis. In this case, the term CRISISLEX_CRISISLEXREC references the lexicon resource and CRISISLEX_C07_SAFETY references the theme "Safety and Security". This theme is associated with the definition "Protection of people/property against harm such as violence or theft" [13].

An area of interest for future work is to understand the terms in more depth by analyzing their database of origin and their context. For example, all terms that begin with "WB" are from the *World Bank Group Topical Taxonomy* database. This will require confirmation from the GDELT developers on the databases that provided each term of interest.

**Network Distance.** The network distance results of English-Russian word pairs associated with the cybersecurity-related terms can be seen in Figure 3. The median distance score is **5** while the mean distance score is **4.64**. Interestingly, rather than being a discordant word, CYBER_ATTACK is well below our median, with a score of **0.54**. The term CRISISLEX_C07_SAFETY has a negative distance metric. We conclude that these two terms must have especially similar network structures between the English and Russian translations of the term. In other words, these terms are contextualized similarly in English and Russian content. Terms that indicate policy, public sector management, information and communication technologies are above average in discordance. This suggests a fundamental difference in how English and Russian speakers understand the role of government in regulating information and communication technologies. This result could be used by policymakers as a quantitative affirmation that both governments understand the role of the public sector in dealing with cybersecurity issues quite differently.

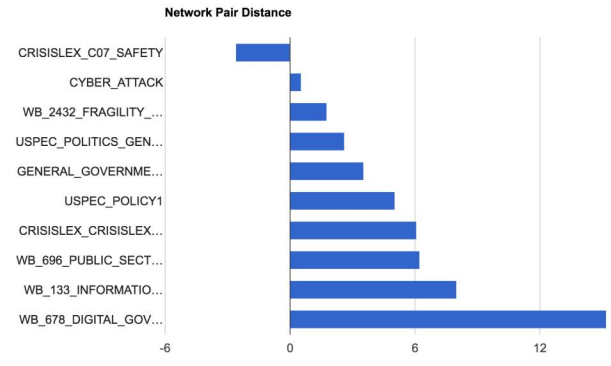



Fig. 3. Network distances of English-Russian Cybersecurity term Network Pairs

**Modularity.** U.S. and Russian policymakers will benefit from further analysis on what exactly contributes to the different viewpoints suggested above. Analyzing community structure within our networks has the potential to inform

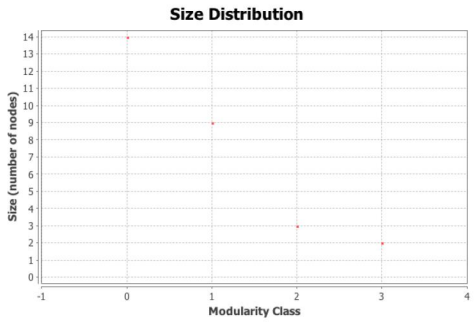



Fig. 4. Community Distribution in English Network WB_678_DIGITAL_GOVERNMENT

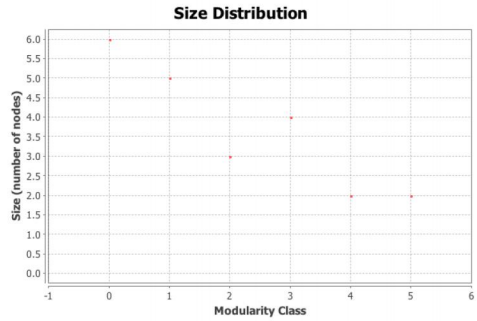



Fig. 5. Community Distribution in Russian Network WB_678_DIGITAL_GOVERNMENT

policymakers on the themes that drive discordance. An area of future work is a detailed analysis of modularity, or community detection, in our most discordant networks. For example, an extension of this work may address the question: "is there a relation between the discordance of a term representing digital government and the subnetwork structure of a cluster representing information and communication technologies?"

In an effort to understand the community structure of our most discordant terms, we begin this work by running a community detection algorithm [14] on our most discordant term, WB_678_DIGITAL_GOVERNMENT. For our parameters, we choose to consider edge weight and set the resolution to **0.5**, where 1 represents an algorithm that prefers large communities and 0 represents an algorithm that prefers small communities. For our English network, the result is 4 communities. The distribution of nodes for each community can be seen in Figure 4. The network with the modularity class indicated by color can be seen in Appendix VII. For our Russian network, the result is 6 communities. The distribution of nodes for each community can be seen in Figure 5. The network with the modularity class indicated by color can be seen in Appendix VIII. A conclusion we draw from our modularity findings is that more themes go into the concept of digital government in Russian content than in English content. In future work, the data collection process should include more node and edge values for highly discordant networks. While even networks with **30** and **27** nodes, respectively, have yielded results, a more robust conclusion can be drawn from community clusters on large networks.

**Network Distance Features.** Finally, we explore our network distance features and hypothesize which feature(s) drive network distance. For each of our signature vectors, we compute the Canberra Distance function of each of our feature values and analyze which features contribute the most to the total distance. From Figure 6, we see that the feature *Average Node Clustering Coefficient* contributed the most while the feature *Number of Neighbors* contributed the least. It is expected that center node features do not contribute much to the distance score because additional graph processing reduced the degree of the center node.

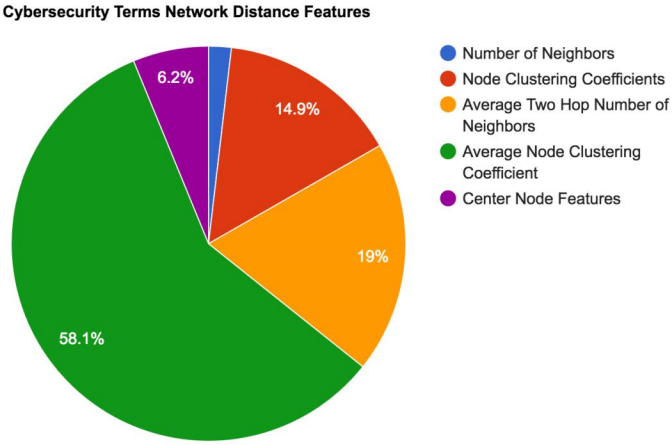



Fig. 6.

The significance of the feature *Average Node Clustering Coefficient* may stem from its interpretation as a global clustering coefficient. A graph is considered small-world if its average clustering coefficient is significantly higher than that of a random graph. We hypothesize that a high average clustering coefficient and a high average two hop number of neighbors metric indicate a network is a small-world network, meaning that all nodes have a low degree of separation. This suggests that between English and Russian translations of a term, one network is highly connected while the other is not. We hypothesize that this is because a word may either be highly contextualized in content, with many links to other words, or not contextualized very much. In other words, either a word means something very specific to an English or Russian speaker, or it is not well-defined. A possible recommendation to policymakers is that when dealing with discordant terms, an effort to understand the exact definition of the term from both perspectives at the start of negotiations will help avoid confusion or conflict escalation later. Future work in this area involves expanding our signature vector to include more features. It would be interesting to note whether the feature *Average Node Clustering Coefficient* still drives network distance or if network distance contributions become more dispersed among the features.

## V. CONCLUSIONS

While both English and Russian speakers contextualize the term cyberattack with the notion of general safety, terms that indicate policy, public sector management, and information security have the largest network distance. We conclude that there is high discordance in how English and Russian speakers understand the role of government in regulating information and communication technologies. From analyzing our network that includes modularity, we conclude that Russian speakers associate the term digital government with more of a focus on media than English speakers do. We also find that more themes go into the Russian contextualization of digital government than in the English contextualization. After exploring our network distance features and computing the feature that most drives network distance, we conclude that terms that are highly discordant represent terms that are highly contextualized in one of the languages. This is problematic for communication because it suggests that either a term is highly nuanced to a speaker and has a very precise definition, or it is not well-defined. We conclude by recommending that policymakers spend the time outlining the exact definition of the discordant terms and agreeing on language parameters before negotiations begin, so as to avoid confusion and increase the chances of compromise.

## VI. ACKNOWLEDGEMENTS

The author would like to thank Dr. Kalev Leetaru for his assistance in the data collection process and the GDELT team for their development of the GDELT Database. The author would like to thank Will Hamilton for his consistent mentorship throughout the development of this paper.

## REFERENCES

[1] Rainie, L., Anderson, J., & Connolly, J. (2014). Cyber attacks likely to increase. http://www.pewinternet.org/2014/10/29/cyber-attacks-likely-to-increase/

[2] Owens, William A., Kenneth W. Dam, and Herbert S. Lin, eds. Technology, policy, law, and ethics regarding US acquisition and use of cyberattack capabilities. National Academies Press, 2009.

[3] Gaouette N. and Labott E., "US and Russia Now in Unpredictable Confrontation." CNN. Accessed October 27, 2016. http://www.cnn.com/2016/10/12/politics/us-russia-tensions-cold-war/.

[4] Lin, H. (2016). Attribution of Malicious Cyber Incidents: From Soup to Nuts

[5] Wheeler, David A., and Gregory N. Larsen. Techniques for cyber attack attribution. No. IDA-P-3792. INSTITUTE FOR DEFENSE ANALYSES ALEXANDRIA VA, 2003. Harvard

[6] Hathaway, Oona A., Rebecca Crootof, Philip Levitz, Haley Nix, Aileen Nowlan, William Perdue, and Julia Spiegel. "The law of cyber-attack." California Law Review (2012): 817-885.

[7] http://www.gdeltproject.org/

[8] Edmonds, Philip. "Choosing the word most typical in context using a lexical co-occurrence network." In Proceedings of the 35th annual meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pp. 507-509. Association for Computational Linguistics, 1997.

[9] Cohen, Aaron M., William R. Hersh, Christopher Dubay, and Kent Spackman. "Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts." BMC bioinformatics 6, no. 1 (2005): 1.

[10] Barbern, Albert, Scott T. Bates, Emilio O. Casamayor, and Noah Fierer. "Using network analysis to explore co-occurrence patterns in soil microbial communities." The ISME journal 6, no. 2 (2012): 343-351. Harvard

[11] Berlingerio, Michele, Danai Koutra, Tina Eliassi-Rad, and Christos Faloutsos. "NetSimile: a scalable approach to size-independent network similarity." arXiv preprint arXiv:1209.2684 (2012).

[12] http://blog.gdeltproject.org/google-bigquery-gkg-2-0-sample-queries/

[13] A. Olteanu, C. Castillo, F. Diaz, S. Vieweg. CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM'14). AAAI Press, Ann Arbor, MI, USA, 2014 Browse on GitHub CrisisLexLexicon-v1.0.zip (2.9 KB)

[14] 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

## APPENDIX I
## CODE

Implementation can be found at `https://github.com/ashemag/CybersecurityContextUSRussia`

## APPENDIX II
## QUERY FOR CYBERSECURITY TERMS

```
SELECT theme, COUNT(*) as count
FROM (
select UNIQUE(REGEXP_REPLACE
(SPLIT(V2Themes,';'), r',.*', '')) theme
from [gdelt-bq:gdeltv2.gkg]
where DATE>20151027000000 and
DATE < 20161026000000
and V2Themes like '%CYBER_ATTACK%')
group by theme
ORDER BY 2 DESC
LIMIT 10
```

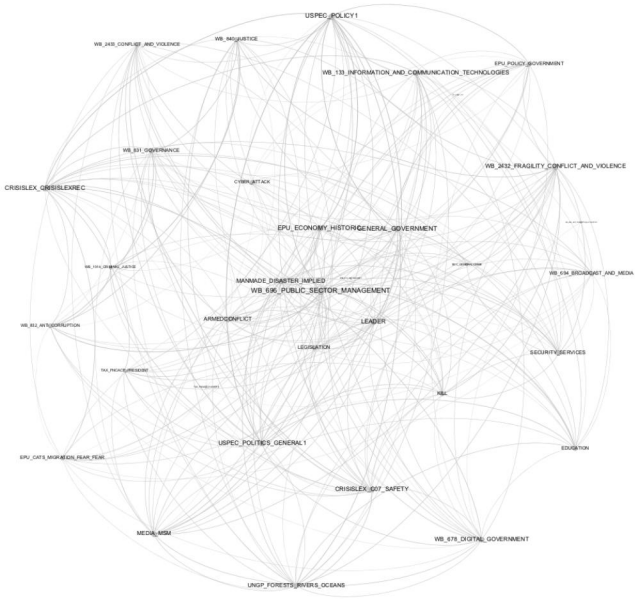
## APPENDIX III
## QUERY FOR CO-OCCURRENCE NETWORK FROM RUSSIAN CONTENT

```
SELECT a.name, b.name, COUNT(*) as count
FROM (FLATTEN(
SELECT GKGRECORDID, UNIQUE(REGEXP_REPLACE
(SPLIT(V2Themes,';'), r',.*', '')) name
FROM [gdelt-bq:gdeltv2.gkg]
WHERE DATE>20151101000000 and
DATE < 20161101000000
and V2Themes like '%CYBER_ATTACK%'
and TranslationInfo like '%srclc:rus%'
,name)) a
JOIN EACH (
SELECT GKGRECORDID, UNIQUE(REGEXP_REPLACE
(SPLIT(V2Themes,';'), r',.*', '')) name
FROM [gdelt-bq:gdeltv2.gkg]
WHERE DATE>20151101000000 and
DATE < 20161101000000
and V2Themes like '%CYBER_ATTACK%'
and TranslationInfo like '%srclc:rus%'
) b
ON a.GKGRECORDID=b.GKGRECORDID
WHERE a.name<b.name
GROUP EACH BY 1,2
ORDER BY 3 DESC
```

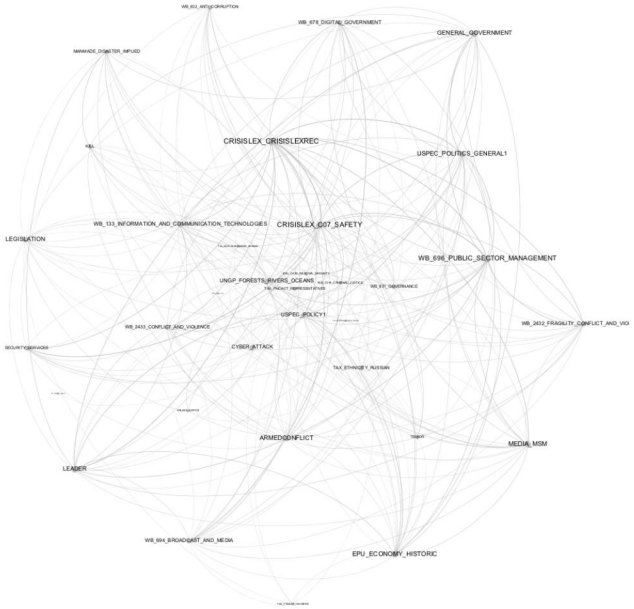
## Appendix IV
## Query for Edge Count Between Two Terms from English Content

```
bq query --format csv "SELECT a.name, b.name
COUNT(*) as count
FROM (FLATTEN(
SELECT GKGRECORDID, UNIQUE(
REGEXP_REPLACE(
SPLIT(V2Themes,';'), r',.*', '')) name
FROM [gdelt-bq:gdeltv2.gkg]
WHERE DATE>20151101000000 and
DATE < 20161101000000 and V2Themes like '%s'
,name)) a
JOIN EACH (
SELECT GKGRECORDID, UNIQUE(
REGEXP_REPLACE(
SPLIT(V2Themes,';'), r',.*', '')) name
FROM [gdelt-bq:gdeltv2.gkg]
WHERE DATE>20151101000000 and
DATE < 20161101000000 and V2Themes like '%s'
) b
ON a.GKGRECORDID=b.GKGRECORDID
WHERE a.name<b.name
GROUP EACH BY 1,2
ORDER BY 3 DESC
LIMIT 1
```

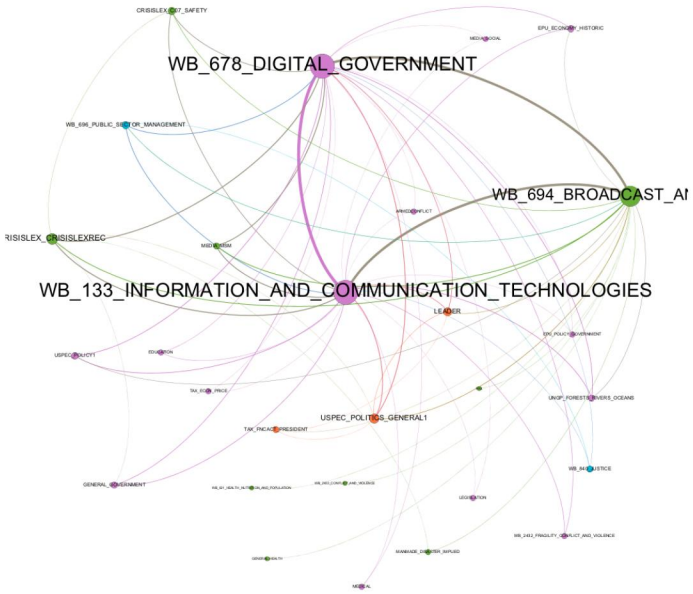
## Appendix V
## Final Network for English Term cyber_attack




## Appendix VI
## Final Network for Russian Term cyber_attack




## Appendix VII
## Community Detection Network for English term WB_678_DIGITAL_GOVERNMENT



# APPENDIX VIII
# COMMUNITY DETECTION NETWORK FOR RUSSIAN TERM WB_678_DIGITAL_GOVERNMENT