CS224W Project Milestone (Fall Quarter 2017): Uncover and Predict Terrorism Network Dynamics

Julia Alison*
jalison@stanford.edu
Stanford University
Stanford, CA

Li Deng*
dengl11@stanford.edu
Stanford University
Stanford, CA

Zheqing (Bill) Zhu *
zheqzhu@stanford.edu
Stanford University
Stanford, CA

1. Introduction

As international concern about the omnipresent terrorist threat mounts, academics and governments are seeking to better understand the underlying dynamics of the organizations behind the attacks. In formulating counter-terrorism strategies, researchers have sought to identify and understand the hidden global terrorism network by applying various network analysis techniques.

As we describe below in our literature review, most of the analysis of terrorist groups has applied social network analysis to known organizations and performed empirical studies following major attacks. We have sought to approach the problem by investigating the dynamics of terrorist organizations in relation to one another as well as to develop an algorithm for predicting and preventing future attacks.

Section 2 includes a review of the relevant literature applying network analysis to understanding terrorist organizations as well as literature on some of the algorithms that we will employ in our study. Section 3 describes the Global Terrorism Database, our main dataset for the project, as well as other datasets that will supplement our work. Section 4 describes the methodology we have employed so far, including the theoretical background for our methods and some pseudocode for our implementation. Finally, Section 5 includes a discussion of our plans for future work on the project.

2. Literature Review

2.1. Theoretical Analysis

Everton (2009) [1], Clauset et al. (2008) [2], and Moon et al. (2007) [3] offer three distinct approaches to the theoretical study of terrorist network dynamics.

Everton (2009) criticizes the significant application of social network analysis to the study of terrorist networks by arguing that the focus on identifying and targeting key players within the network is misplaced. He cites previous work that demonstrated the effectiveness of taking out hubs and cliques within otherwise sparse networks and argues that future work should be directed toward understanding the overall topology of terrorist networks. Citing various examples at both a personal and firm level, he argues that organizations and social groups are most effective with an optimal mix of strong and weak ties. Understanding the specific hierarchical structure within a terrorist organization (centralized leadership, decentralized, etc.) has strategic implications, as it can help security forces target specific actors in a centralized network to disrupt its effectiveness. These measurements include identification of leadership as highly connected nodes within the network as well as an understanding of the cosmopolitanism of the network via measurements of the local clustering coefficients and average path lengths. However, his explanation is incomplete, as connections within networks are often unknown (so his empirical study of legal organizations like firms is somewhat unrelated). However, once a network has been inferred, the measures he's cited can be helpful for understanding the overall landscape of the terrorist organization.

Clauset et al. (2008) further supplemented our understanding of the underlying network structure of terror cells by building a hierarchal structure of a terrorist network via a maximum likelihood model to infer connections between nodes. They compared this model to a true terrorist network and found the two networks similar along metrics like average clustering coefficient and SCC size.

Moon et al.'s (2007) approach is more technical in nature. Using a network of terrorist communication gathered after the 1998 US Embassy bombing in Kenya, the authors built a meta-structure of tasks and the agents assigned to complete them in order to infer the players who would have had contact in order to carry out the attack. The researchers were able to establish connections within the covert terrorist network by finding the shortest path between the individual assigned and the person with the necessary expertise

*Joint first author

or information. The paper then applies social network analysis techniques by calculating betweenness centrality and total degree centrality on nodes (labeled as particular figures in the organization) for each of the models they built to identify key players within each theoretical organizational structure. The paper's primary strength is in its innovative approach to understanding why edges between actors within the terrorist network exist. However, its methods for studying those networks are not particularly rigorous, and it does not provide a metric for testing any of the constructed models. While they apply the study to a single terrorist cell, the approach of building the network via spread of information and methods used seems promising for our development of a global terrorist organizational structure.

2.2. Case Study

In addition to theoretical analyses that provide foundations to depict the underlying structure of terrorism networks, researchers have conducted case studies that provide insights in individual terrorist organizations. Belli et al. [4] provided results from a case study that explores the social network of internal members of the "Hammound Enterprise", which was involved in trade diversion in order to finance terrorism organizations in Michigan. They found out that in these groups, members are highly interconnected, making organizations more efficient while vulnerable to detection. With key player analysis, three ringleaders are detected with a few secondary leaders, most of whom are Islamist extremists, which shows the property of an idea-centric organization. These analyses are very representative across many case studies we have found.

Krebs 2002 [5] is widely referenced in literature about terrorist network analysis. Following the September 11 attacks, the author used publicly available information about the 19 hijackers to construct a network of weak and strong ties based on the nature of their relationships with one another. By computing the degree distribution, betweenness, and closeness of the nodes, this paper depicts a sketch of the covert network behind the scenes, and allows him to identify the clear leader among the hijackers. Also from the analysis on clustering coefficient(0.4), and average path length among the nodes(4.75), the authors found out that this covert network "trade efficiency for secrecy" by remaining quite small and operating with little outside assistance.

Krebs' paper is one of the foundational documents in applying network analysis to terrorist organizations post-9/11. Because his research was conducted after the fact on a well-publicized case, he is able to delve into the nature of connections among actors in considerable depth, which lends credence to his findings about this specific group. However, his work does not lend itself to application on the larger terrorist landscape, because in most cases, information about the trust and familial connections among actors in a covert network is difficult to come by and verify. Thus, the insights gained from particular case studies can inform some of our approach to larger networks (i.e. identifying key players based on centrality, looking for hubs in a sparse network, etc.), but their approach is limited in scope and application to future study.

Instead of exploring the internal structures of each group, a global picture of the interaction between different groups may provide crucial linkage that connects pieces of small cluster networks. For instance, if we can establish a concrete theory of the collaboration between ISIS and extremists groups within the U.S., we can potentially figure out many sources of resources of terrorist attacks in the U.S. Another benefit of studying the global group network of terrorism is that it helps understand the global shift of terrorism distribution and predict future moves.

Aside from global network of terrorist groups, another useful study is to clarify the connection between terrorist events and these groups. Events can also be formed as networks. For instance, a table of events with date, location and group can be treated as nodes and linked by data like transportation network to study how resources and manpower are moved within these networks. When the data is linked by social networks like Twitter or Facebook, we can hope to understand how these extreme ideas are spread through the online community and how people are recruited to these groups.

Finally, a key aspect of terrorism is time. If we can establish a time dependent trend of terrorism activities, these analyses would be at great help building theories that effectively limit the expansion of terrorism.

We believe all three aspects above can form research with great impact in the counter-terrorism network research community.

2.3. Methodologies

Given that previous studies of terrorist organizations are either theoretically simplistic or limited in scope to single cells with well-documented communication, our project seeks to apply cutting-edge algorithms from other aspects of network analysis to the global terrorist network.

Kim and Leskovec [6] cast the network completion problem, which aims to infer missing edges and nodes in a network, to a Kronecker's Expectation Maximization algorithm. Using their method, one can infer or uncover the hidden network in a terrorism network based current event network. The paper outlines a methodology for inferring a network when a significant number of nodes are missing or unobserved. However, while individual players within terrorist organizations tend to be unknown, we believe that the global terrorist network's nodes (terrorist organizations) are well understood, but edges (connections between organiza-

tions) tend to be much less clear.

Gomez-Rodriguez et al. (2011) [7] propose a NetRate algorithm that inters the spatiotemporal dynamics that generate observed infections. The algorithm addresses three key models in network dynamics including transmission likelihood, probability of survival given a cascade and network inference problem. This algorithm would help understand the temporal and geographical dynamics of terrorism networks. This algorithm will contribute to our understanding of the spread of terrorism across regions, which can direct counterterrorism efforts toward stopping it from taking hold of a particular group. NetRate is more applicable to information thought to be viral (i.e. the spread of particular internet memes), so our application of this methodology to the spread of terrorism may be limited, but it will provide an additional predictive measure to infer the next terrorist attack. The algorithm requires 5,000 cascades in their evaluation in order to achieve mean absolute error values below 20 percent, so the accuracy of the model for our dataset of terrorist events (with a limited number of documented cascade events within particular regions) is dubious.

However, we build on this inference of edges between nodes by applying and comparing three algorithms for edge inference. We begin with the simplest implementation by Alpay et al. (2011) [8], an algorithm called FastInf. Then, we employ the more computationally rigorous NetInf method from Gomez Rodriguez et al. (2012) [9], which infers static edges between nodes based on observed cascades of information between the network. Finally, we apply the InfoPath cascade algorithm, described in Gomez-Rodriguez, Leskovec, and Schlkopf (2013) [10], which includes the dynamic component of network edges forming and changing over time. These algorithms are described in more detail in the methodology section.

Buhur et al. (2015) [11], although a different target than ours, examined the underlying network for criminals with some interesting approaches. They explored the criminal network within three scopes: link prediction, hidden link recovery and network breakdown. They discovered a social-network-like pattern in link prediction where hidden links are more likely given a shorter path length. Moreover, they established a solid attach algorithm with iterative pagerank that achieved great outcome in breaking down weakly connected components.

3. Dataset

3.1. Global Terrorism Database

Global Terrorism Database(GTD)[12] is an open-source database including information on terrorist events around the world from 1970 through 2016. It contains information about 170350 terrorist events, and a total of 135 attributes for each event, including exact date, position, group, weapon, casualty and so on.

Here is the geographic distribution of events recorded by GTD around the globe:

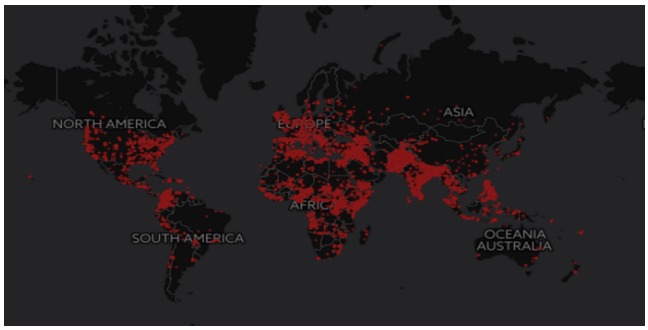


Figure 1: Geographic Distribution from GTD (Source: Kaggle)

Our first step is to understand the data. Here are the top 10 groups by their number of terrorist events:

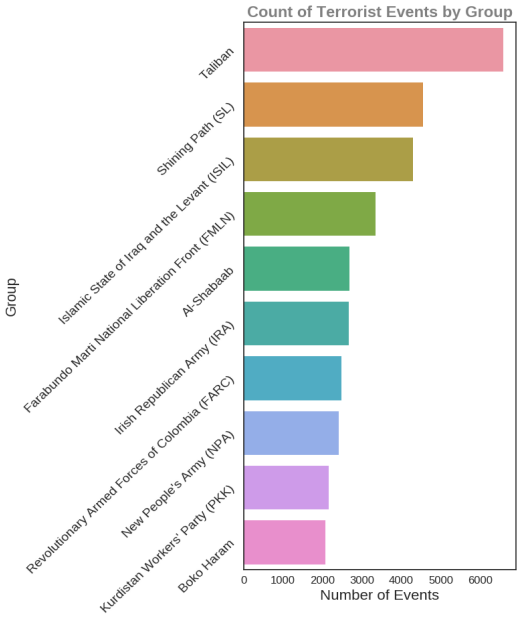


Figure 2: Count of Events by Terrorist Group

Here are the top 10 groups by country:

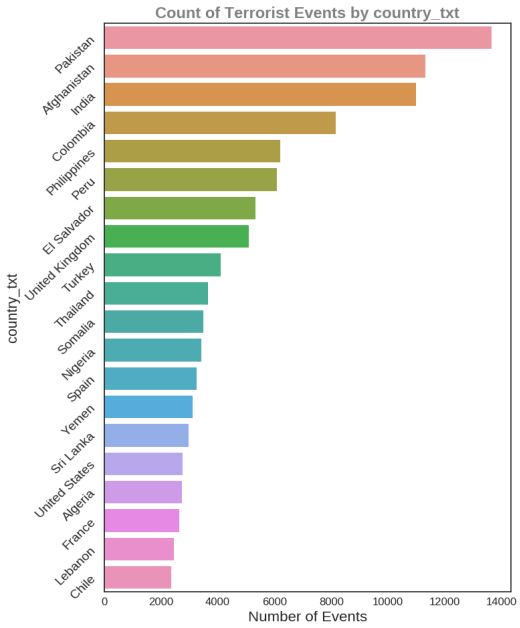


Figure 3: Count of Events by Country

3.2. Open Flight Database

In addition to GTD, we here introduce Open Flight Database (https://openflights.org/data.html), which is a open-source database that maintains a network of over 6,000 global airline connections. We believe that in order to better understand the connection between events and geolocations, geographical connectivity is essential to our success in analyzing geographical dynamics of the terrorism event network.

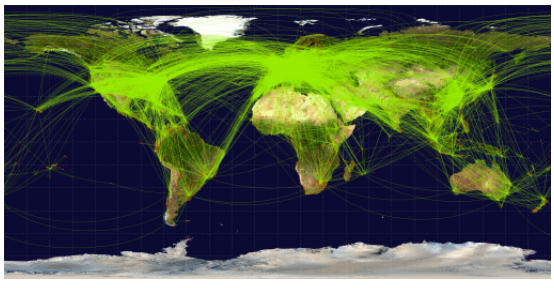


Figure 4: Open Flight Database (Source: openflights.org)

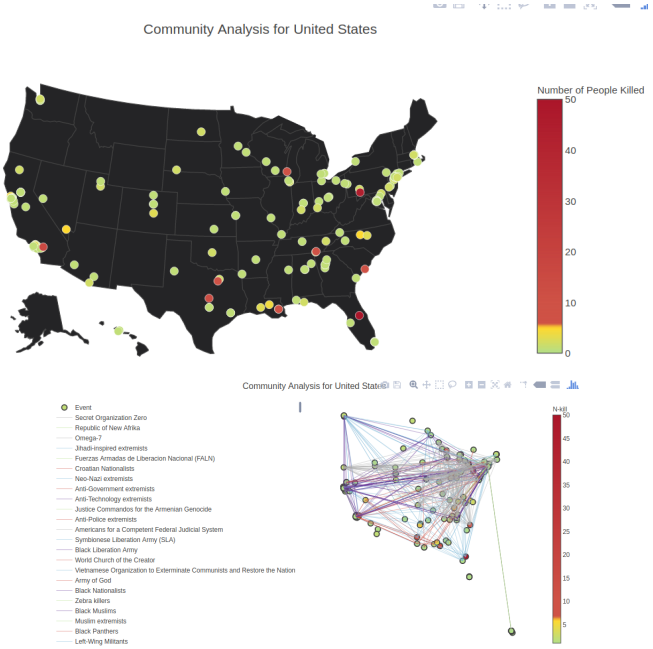
The idea to link terrorism events with transportation network is to extract additional graphical images that can be used to make predictions about location specific information. We assume here that airline connections indicate communicative and transportational connection between two locations and will present clusters depending on geographical data.

4. Methods, Algorithms and Evaluation

4.1. Link Construction

Each row in the GTD dataset is a terrorist event. Our first approach to construct a graph out of it was to create links among events who shared some attribute.

One of the important attributes is the attack type, since its reveals the motivation of the terrorist for a certain event. And it's reasonable that terrorist groups who share the same motivation may corroborate with each other. So we connect the events together if they have the same attack type, as shown in 4.1. But one disadvantage of this approach is that we end up having cliques of fully connected components.



4.2. Temporal Evolution

Since Taliban is the largest terrorist group by the number of events, we pick it as the case study to understand its evolution or cascading pattern in the timeline.

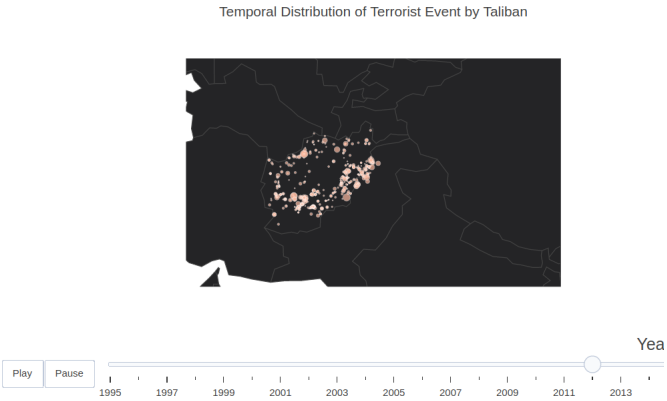


Figure 5: Temporal Evolution of Taliban Terrorist Events

Based on this animation tool, our next step is to get intuition from the visualization, and analyze its cascading pattern numerically.

4.3. Relation Network Analysis

The GTD dataset contains a "relation" column, which is a list of Event IDs of related terrorist events. This can be of huge help for constructing the relationship network among the events.

We first created the visualization of global relation network as shown in 6

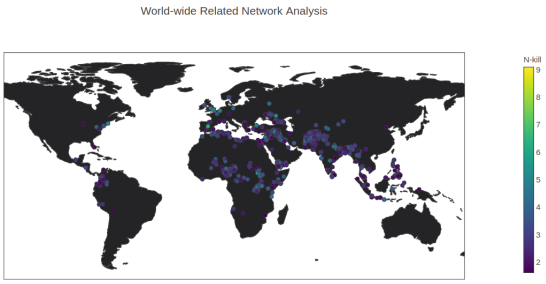


Figure 6: World-wide Relation Network

And then we focus on the relation network of the United States. The USA relation network contains 463 nodes as shown in 7, and 883 edges.

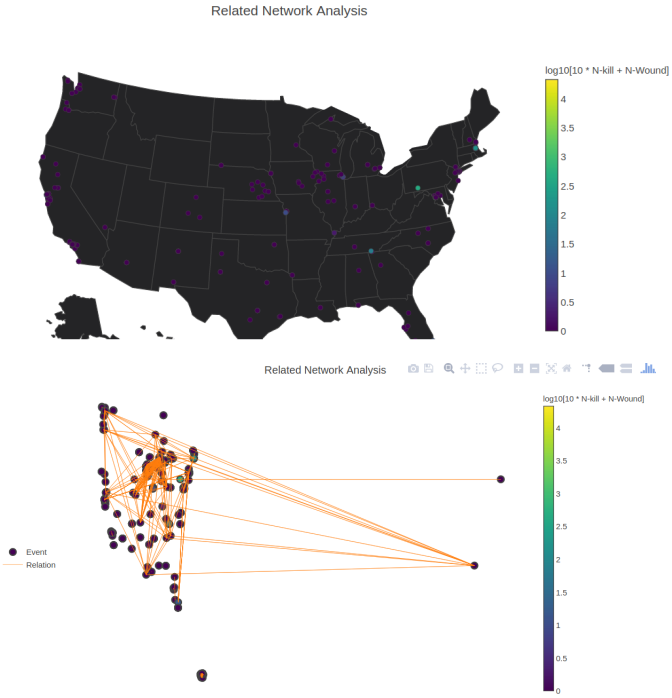


Figure 7: USA Relation Network

- **Degree Distribution**

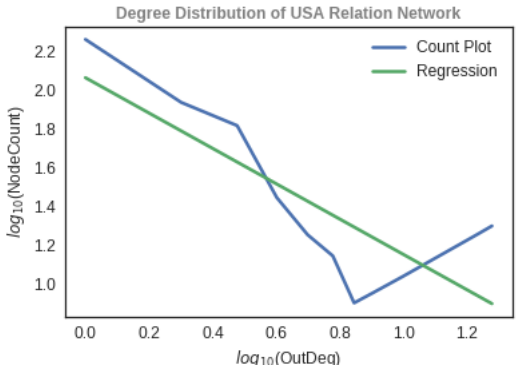


Figure 8: Degree Distribution and Regression of USA Relation Network

We also try to understand the correlation between node degrees and the severity of the event corresponding to the node in relation network. The top 100 events(nodes in the relation network) were extracted, with their degrees compared against with their consequences. We may have different measure for the severity(impact of consequence) of a terrorist event. Here we consider three possible measures: $nKill$(number of people killed), $nWound$(number of people injured), and $10 \cdot nKill + nWound$(a scaled combination of the two), as shown in 9.

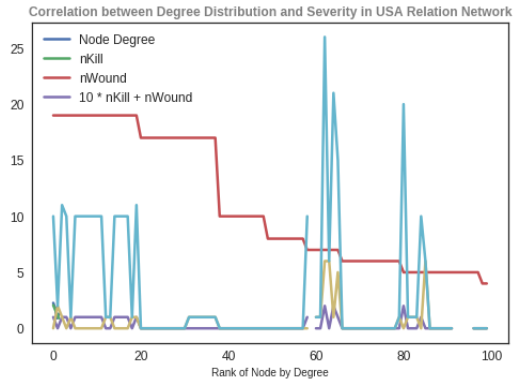


Figure 9: Correlation between Degree Distribution and Severity of USA Relation Network

The correlation turns out to be quite interesting: when nodes have very high degrees(far left size of the figure), they have high severity, which is expected, since events highly influential should have higher potential to cause greater kills and injuries. While on the far right size of the figure, events have relative low degrees, many of them may still have great severity, i.e. causing lots of kills or injuries. We will look into the reason behind it in the next phase of our project, and we expect those events with high severity but low degrees more likely caused by individual terrorists, while thoese events with high severity and high degrees more likely to be caused by groups.

- **Weakly Connected Components**: 152 Weakly Connected Components
- **Max Weakly Connected Component**: the largest WCC in the USA relation network contains 20 nodes, and 190 edges

4.4. Terrorist Group Prediction

Prediction for terrorist events would be very useful in reality, but difficult in implementation. So instead of predicting the occurrences of new events, we limit our scope to prediction of perpetrator group of an event that just happened. There are more than 600 terrorist groups in total, which is unreliable to predict. So we only consider a subdataset from GTD that are caused by the top 10 groups, according to 2. We pick 10 of the features from existing GTD dataset, and predict a new event to be one of the 10 groups. We decide to use XGBoost as the classification algorithm, due to its outstanding performance and accuracy in recent data competitions. Surprisingly, we found that the predict for 10 groups were almost perfect, training and prediction accuracy were both around 99%. With further investigation, we realized this is because there is a very strong relationship between the group and the country where the event happened, i.e. usually by just where the event happened, we could be table to tell who(which group) caused it. So we removed "Country" from the feature list, and both training and prediction accuracy dropped to around 60%. Our next step is to extract more signals from our network analysis, to aid the prediction in an more informed way. So instead of just using existing information from GTD dataset, we could validate our analysis of network by applying it to the prediction problem.

4.5. Understanding Links Between Global Terrorist Organization Nodes

4.5.1 Hand-Labeled Data

As described above, the GTD includes a column for each attack of the IDs of "related" attacks. We used the existence of "related" attacks to hand-label the relationships between different terror organizations with the following algorithm: For each attack in the database, we identify the perpetrating group and then create links between that group and all groups in the database carrying out related attacks. This yields a multigraph with edges between nodes for all related attack tuples that they share. We omit self-edges from our analysis.

This model yields an extremely sparse graph wherein the majority of edges are self-edges. We will work toward deriving a useful result from this implementation, but at this point we are seeking an alternative measure of group-relatedness.

This approach is the most simplistic approach to inferring connections between organizations, because it relies on the public acknowledgement of the relationship between two attacks. The most significant force driving that acknowledgement is the attacks sharing a main perpetrator, which does not contribute to our understanding of the relationships between organizations. However, we will use this observable collaboration network as a baseline for our edge inference models described below.

4.5.2 FastInf and NetInf

The database has information about the target, style of attack, and weapon used to carry out the attack, as well as its location. We used these features to label cascades of similar attacks temporally, which allowed us to implement the NetInf algorithm to infer the most probable edges between terrorist groups in the graph.

Because this implementation takes a significant amount of time, for our milestone we limited the study to terrorist attacks in the Middle East since 1996. Under the assumption that a previous attack in a similar style might have influenced a future attack, we consider that any earlier attack may have influenced the later attack. In the simplest implementation, we assume that each prior attack had an equal probability of causing the infection. We then implement

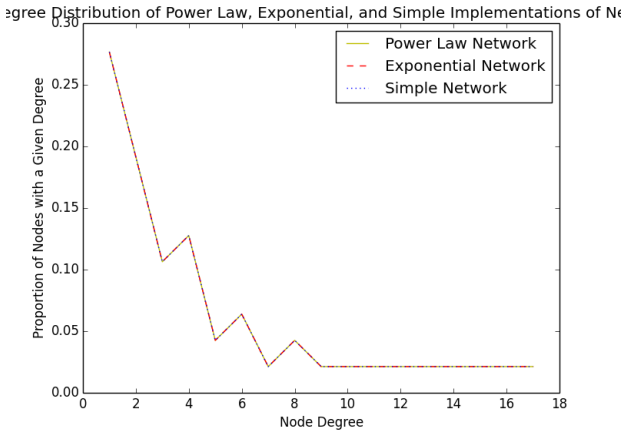
edge weights with exponential and power law transmission models by calculating edge weights as follows: Exponential Transmission:

$$w_c(i,j) = \frac{e^{time_j - time_i}}{\sum_{n_{i'}, t_{i'} < t_j} e^{time_j - time_i}}$$

Power-Law Transmission:

$$w_c(i,j) = \frac{(t_j - t_i)}{\sum_{n_{i'}, t_{i'} < t_j} (t_j - t_{i'})})$$

The degree distributions are graphed belowinterestingly, for our small network, each model of spread yields an identical graph of connections. The graph is still extremely sparse, and most nodes are considered unconnected to any other group. However, the connections we can establish will be a feature to consider in trying to predict future attacks via our particle filter implementation, as we can infer that copycat attacks can spur future ones.



4.5.3 NetRate and InfoPath

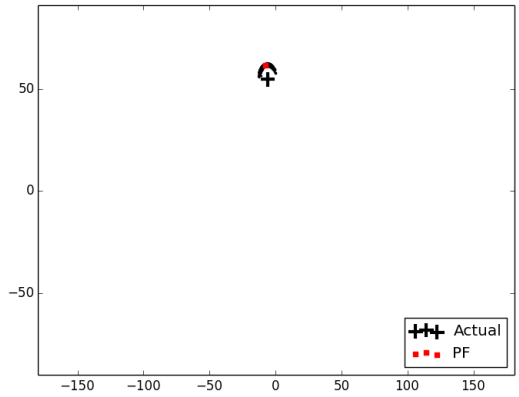
As we move to the final project, we plan to implement the InfoPath algorithm from Gomez-Rodriguez et al. (2011), which takes in cascades as FastInf does, but accounts for the dynamic nature of the nodes changing over time. The model allows connections over time to expire, which makes sense given the 50 year time horizon of our dataset. We can compare the graph of connections yielded by this approach to the graph yielded by the more simplistic algorithm.

5. Terrorism Event Prediction

Terrorism event prediction and location forecasting is an essential topic when analyzing terrorist event networks. Since most events occur on different dates, in order to make reasonable predictions with enough data, we discretize the event space across time into slices of events in each year. Given that data, we use the following two approaches to make forecasts of terrorist events regarding group, city and country.

5.1. Particle Filter Based Event Tracking

One intuitive understanding of terrorism is that they spread. The typical way for terrorism to spread is that first the terrorist group would recruit radical people from their target locations and brainwash them. Then they start to use these people to either make terrorist attacks or recruit more people. From the procedure, we can observe that the spread is directional, which suggests that the spread of event could be tracked by a motion tracking system. In our approach, we adapt particle filter to model the movement of several terrorist groups to test our hypothesis of directionality.



The particle filter generally gives a scattered sample of weighted particle and prediction like above. this is shown as an example an instance particle filter prediction. In the figure above, the x axis is latitude and the y axis is longitude. We use particle filter to track a terrorist group by event location at each occurrence. What particle filter does is as the following: first, the particle filter initializes uniform distribution of particles all around the world (the latitude, longitude grid). Each time when an event from the same group occurs, the algorithm updates the weight of each particle. Finally, when the algorithm is called to predict the next event, the particle filter calculates the weighted expected particle from all particles scattered. Particle filter is specifically used here due to the suspected non-linearity and non-Gaussian attributes of terrorism movements.[13]

Using particle filters provides us with an average of 532.37 miles in Vincenty distance for predictions of location of Irish Republican Army events and 688.93 miles in Vincenty distance for ISIS events.

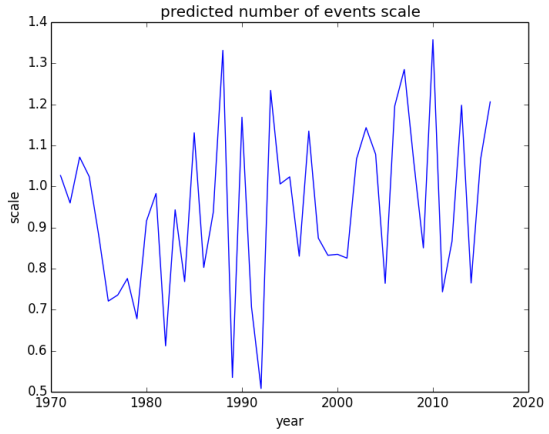
5.2. SIS Model for Terrorism

Apart from a motion tracking perspective, from the network perspective, one intuitive way of connecting the event network is through traffic network. As a preliminary step, we connect all cities that ever have terrorism events with airline route data from openflight. Using time sliced network, we treat terrorism as an epidemic and follows an SIS

model.

The SIS model parameters are estimated as the following, where β accounts for the infection rate and δ accounts for the probability of coming back to susceptible, $\beta = \frac{N_{inf,nbr,sus}}{N_{sus}}$ and $\delta = \frac{N_{inf',sus}}{N_{inf}}$, where inf means node is infected, nbr means having a neighbor with attack, sus means susceptible and inf' means was attacked in the last time slice.

Using the model specified above, we are able to get the following result for predicting the number of terrorist events per time window.



The above figure shows the ratio between predicted number of events and the actual number of events based on each year-slice of data.

6. Moving Forward

Based on the visualizations and animation tools we developed, and our experiments so far, we have had a solid infrastructure for further analysis, and better understanding of the data, as well as the problems.

Our ultimate goal for this project is to refine our ability to predict terrorist attacks using the both the particle filter and information about the features of the attacks. Understanding the attacks as cascades allows us to further understand connections between groups and will contribute to our understanding of future attacks as we tune cascade parameters based on known attacks and then predict the additional geographic infections within the cascade. Algorithms we have implemented like FastInf and algorithms we are developing like InfoPath will supplement this work.

The particle filter uses geographic information, and we can further refine our understanding of predictable spreads of nodes via information about roadways and traffic between possible attack sites and the number of attacks by more advanced outbreak detection

Our priority in the coming weeks is to extend our use of the algorithms discussed above to the full dataset, and to justify parameterizations of the data when appropriate (i.e. predicting US-specific attacks).

References

[1] F. Everton, Sean. Network topography, key players and terrorist networks, 2009.

[2] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 2008.

[3] Alexander H. Levis Il-Chul Moon, Kathleen M. Carley. Vulnerability assessment on adversarial organization: Unifying command and control structure analysis and social network analysis. *EEE Intelligent Systems, Special issue on Special issue on Social Computing*, 22(5):40–49, 2007.

[4] Roberta Belli, Joshua D. Freilich, Steven M. Chermak, and Katherine Boyd. Exploring the u.s. crime-terror nexus: Terrorist networks and trade diversion. *National Consortium for the Study of Terrorism and Responses to Terrorism*, 2014.

[5] Valdis Krebs. Uncloaking terrorist networks. *First Monday*, 7(4), 2002.

[6] Myunghwan Kim and Jure Leskovec. The network completion problem: Inferring missing nodes and edges in networks. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 47–58. SIAM, 2011.

[7] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *In Proceedings of the 28 th International Conference on Machine Learning*, 2011.

[8] Altan Alpay, Deniz Demir, and Jie Yang. Fastinf: A fast algorithm to infer social networks from cascades. *CS 224W Final Project*, 2011.

[9] Manuel Gomez-Rodriguez, Jure Leskovec, and Andres Krause. Inferring networks of diffusion and influence. *ACM Trans. Knowl. Discov. Data*, 5(21), 2012.

[10] Manuel Gomez-Rodriguez, Jure Leskovec, and Bernhard Schlkopf. Structure and dynamics of information pathways in online media. *WSDM*, 2013.

[11] Emrah Budur, Seungmin Lee, and Vein S Kong. Structural analysis of criminal network and predicting hidden links using machine learning. 2015.

[12] Gary LaFree and Laura Dugan. Introducing the global terrorism database. *Terrorism and Political Violence*, 19(2):181–204, 2007.

[13] Pierre Del Moral. Non-linear filtering: interacting particle resolution. *Markov processes and related fields*, 2(4):555–581, 1996.