Generating Fraud: Agent Based Financial Network Modeling

Matthew Koehler, Brian Tivnan, and Eric Bloedorn

The MITRE Corporation
7515 Colshire Dr.
McLean, VA 22102-7508
mkoehler, btivnan, and bloedorn@mitre.org

Abstract

In this paper, we describe research and an application of agent based modeling to create financial network data. Creating a dataset of this type presented some unique challenges. First, the dataset we are trying to emulate is large and sparsely connected (20 million nodes, 20 million edges, in 500GB). Second, it includes multiple types of entities and relationships. A system made up of multiple types of entities with various relationships is tailor made for agent based modeling. Third, this dataset is being created as part of a larger project that is creating graph analysis tools that will work with massive, dynamic datasets. Therefore, it is important that we be able to control what the generated dataset contains so we can test various parts of our graph analysis system. An initial agent based model has been created using Netlogo. This prototype is being created iteratively as we continue to investigate the patterns and other features within the actual dataset. The domain in which the graph analysis tools are to be used is, understandably, of a sensitive nature. We wish to keep the datasets we produce unclassified so they can be released to the academic and analytic communities to aid in collaboration. This presents its own challenges as we need to produce a dataset that is a reasonable facsimile of the actual data for meaningful collaboration, however not so similar as to represent any sort of unreasonable disclosure of information.

Contact:

Matthew Koehler

The MITRE Corporation, H305

7515 Colshire Dr., McLean, VA 22102

TEL: (703) 983-1214

FAX: (703) 983-1379

mkoehler@mitre.org

Keywords: Agent Based Modeling, Financial Networks, NetLogo, Data Creation

Generating Fraud: Agent Based Financial Network Modeling

Matthew Koehler, Brian Tivnan, and Eric Bloedorn

1. Introduction

The MITRE Corporation is a set of federally funded research and development centers. One of the projects taken on by MITRE is work with various US Federal government organizations that look at and investigate financial networks and crimes. Understanding financial networks are very important to combating, *inter alia*, tax evasion, organized crime, drug trafficking, and terrorist financing. As an indication of the importance the US Federal government places on this, please see Section 602 of the US PATRIOT Act [U.S. PATRIOT ACT 2005]. One of the features of this type of investigation is the staggering size and dynamism of the dataset associated with the movement of money and payment of taxes even if one limits the sample to transactions within the US. This paper describes work done as part of a larger effort to create tools that can overcome the analytic challenges of working with data of this type, see [Bloedorn 2005] for a complete description of the tools developed for massive scale relational graph analysis.

1.1 Constraints and Requirements

The most obvious constraint is simply the size of the dataset. Even a small sample of it can be huge (approximately 20 million records), capturing over 20 million nodes and 20 million edges, resulting in 500GB of storage across numerous database tables. Furthermore, the data grows constantly over time as individuals and businesses file tax returns and banks move money all over the world; therefore, static snapshots are simply insufficient [Bloedorn 2005]. An additional requirement is the ability to handle multiple types of entities and relationships within the dataset as well as within subsequent analyses. Moreover, the datasets come from sensitive sources and contain information that cannot be released to the general public, making outside collaboration difficult. Finally, we needed the ability to create datasets with particular network structures and other features so that we might test the relational graph analysis tools being created in a more structured manner.

Given these constraints it was decided that we needed to create a dataset that would emulate sufficiently the real data so as to adequately represent the challenges associated with it, but would be releasable to the academic and greater analytic community. Given the pervasive heterogeneity of the dataset and its importance to the analyses to be preformed upon the data, and the dynamics associated with the dataset in general, we chose to use an agent based framework for data creation. Although it could be argued that an agent based model is not the most efficient way to generate data, we felt that these performance constraints were outweighed by importance of heterogeneity and biases caused by the actions of individual agents embedded within the financial and social networks.

2. The Agent Based Model

As stated above, an agent based framework was chosen for the model structure due to the interdependence between the financial networks and the underlying social networks. Agent based models can be appropriate in these circumstances as they can capture the behavior of social networks and the entities embedded within them. See [Axtell 2000] for a complete discussion on the use of agent based models. The initial purpose of this model was to generate a dataset that has an analogous structure and size to the other datasets we are using. As the modeling proceeded it became clear that this type of simulation tool could serve another use, that of "analyst sandbox." (i.e., organizational flight simulator ([Sterman 1992]; [Hazy & Tivnan 2004])) In this regard, it could allow an analyst to hypothesize about the behavior of agents doing particular illegal transactions and then create a dataset to examine what the "trail" would look like, if in fact, the agents did behave as hypothesized. Furthermore, having a releasable dataset would allow a larger pool of users to create queries and potential enhancements to the tools available to interested analysts. The initial prototype was created with NetLogo [Wilensky 1999]. Netlogo was chosen because of the speed and ease with which one can develop working models.

2.1 Model Overview

This current model is one of transactions. Physical space is not modeled explicitly, nor is agent movement. Agents are created with a set of characteristics representing, *inter alia*, location, nationality, and predisposition to engage in illegal activities. Using these characteristics the agents create a financial transaction network with which they will perform transactions during the course of the model run. The creation of the financial network is biased by the individual

characteristics of the agents. During a model run if an agent decides to perform a financial transaction it can choose to do it legally or illegally. If the transaction is legal the agent will simply move money from its account to an account of the recipient. If the transaction is to be done illegally, the agent decides on a layering scheme (moving the money through multiple accounts) and a structuring amount (breaking the transaction into multiple smaller amounts to avoid some bank reporting requirements). Output from the model includes the entire transaction network of each agent, a summary of each transaction performed by an agent, and all transactions between agents (timestamp, source, sink, and amount).

2.2 Agent Details

In this model agents represent three types of entities: individuals, businesses, and containers. Each of these types was further subdivided. Businesses were broken into for-profit, not-for-profit, shell, and trust. Containers are divided into bank account and hawala (a type of informal money transfer system). Individuals are not broken down any further in the current model and are given the designation: people. Upon agent instantiation all agents set their location and nationality. For the purposes of this prototype, nationality was defined by a random number drawn from a uniform distribution ranging from 0 to 100. Agents also have a state, city, and street. Although location is not explicitly modeled it is used when agents wish to transfer money into accounts in locations that are foreign. All of these values are random integers drawn from an exponential distribution. An exponential distribution was chosen to represent to relative abundance of certain locations in the actual datasets. This distribution has a mean of four for state and city, and a mean of five for street. If the agents had a value greater than 20 for their state, they considered themselves to have a location that is foreign, otherwise they considered themselves domestic.

Finally, agents initialized another set of parameters to create a likelihood to perform illegal activities. These include a predisposition value from 0 to 1. The closer to 1 the more likely the agent is to try to do something illegal in their financial transactions. The final parameter is a boolean variable to designate whether or not the agent is on a “watch list” or not. This is a function of the agent’s location, nationality, and predisposition. To add a bit of noise to this, a small percentage of agents are chosen at random to be placed on the “watch list.” Upon completion of the initialization of the agents, the agents write all their parameter values to a file.

2.3 Agent Network Creation

Next, the social and financial networks are created. The containers’ social networks were simply 60 percent of the other containers not including themselves. The thinking here was simply that most bank accounts can transfer money to most other bank accounts. The individuals and the businesses also created a social network with which they would interact during the course of the run. The initial social structure was created by each agent drawing a random number from an exponential distribution with mean 2. This number was created with a floor of 1 to ensure that everyone was connected to at least one other agent. Businesses and individuals then created links to that number of agents (without duplication). Individuals and businesses then augmented this network with a few specific connections. These included a business and an individual that shared the same state as the networking agent, and a business with type equal to shell and another business with type equal to trust (shells and trust are types of entities that are often used in tax evasion schemes). Further, agents added an individual to their social network that was the same nationality as themselves but had a location of foreign. Agents with a sufficiently high predisposition for illegal activity also added a container to their container list that had a very high predisposition. Finally, individuals and businesses created a list of accounts they may use during the course of the run. Again this was a random number drawn from an exponential distribution with mean 1. This number was created with a floor of 1 and a ceiling of 10. Then, in a new file, the agents write out their social network.

2.4 Model Run Details

When the model is run, one percent of the agents are asked, at each time step, to perform a financial transaction. This is made up of the following steps:

1. First the agents clear any values that may have been set during a previous financial transaction.
2. The agents receive an amount of money with which to perform the transaction.
3. Next the agents pick a destination for the transaction. This is a member of the agent’s social network. On some occasions, however, an agent will pick a random agent with a foreign location and nationality, or pick an agent that has the closest predisposition value to their predisposition value. This was done to introduce some random variance in the transaction network.

4. After picking a destination the agents decide on the structure of the financial transaction. If the initiating agent and the destination both have relatively low predispositions for illegal activities then the initiating agent chooses a basic transaction type with no structuring. If the predispositions of these agents are high but the transaction amount is less than $10,000, then the agent will not structure the transaction but will choose an informal method (hawalas) for the transactions; otherwise the agent decides to structure the transaction but will not yet pick a method.

5. If the agent did not pick a method in the above procedure, it will now pick one. The structures that the agent has available to it are, at this time, fixed. They are loosely based upon structures used for tax evasion and money laundering. The structure is chosen based upon the characteristics of the initiating and the destination agent. These structures included "typical" money laundering and tax fraud schemes involving numerous transactions through shells, trusts, other individuals, and so on designed to make it very difficult to determine the initial source of the money.

6. Once a method is chosen the agent creates a schedule for the transaction. If they are not structuring the transaction the whole amount is transferred from the initiating agent to the target. If, however, the agent is structuring the transaction then the agent creates a schedule of transactions, either uniform or heterogeneous, that are less than $10,000. Finally, the agent builds a timetable for each transaction until the entire amount of the transaction is reached.

In addition to the activities of this one percent of agents, there is some probability that up to four other agents will be told to create very specific tax evasion structures of particular interest.

2.5 "Verifying and Validating" Initial Results

The model has created a number of large data sets and seems to be running correctly. It has been run with approximately 18,000 agents evenly split between individuals, businesses, and containers. All the aforementioned structures are found in the data sets – an initial indication of the model's ontological adequacy (i.e., the model as reflective of the causal dynamics of tax fraud [McKelvey 2002]; [Hazy & Tivnan 2004]). Unfortunately, the resulting transactional network is highly connected, much more highly connected than the actual data indicates. In practical terms what that means is that it is very difficult for our frequent substructure graph analysis tools to find the correct substructures without some corruption of the search by other transactions. For example, when running a query with our tools it will, on occasion, return a correct match on a structure of interest, however, one of the links may be from a different set of transactions as indicated by a different transaction amount.

2.6 Next Steps

Even with 18,000 agents we are under representing the size of the banking sector in the US. We are currently moving the simulation in to Repast [Collier & Howe 2003] so that we can increase its scale. Furthermore, we are developing a methodology to distribute the simulation on a cluster computer. Once again, initial prototyping will be done in NetLogo for speed and simplicity. We currently have the infrastructure in place to run NetLogo models on large scale cluster computers Koehler et al [Koehler, et al. 2004]. However, this is done in a purely parallel manner—the nodes do not communicate any model state information to each other. We are currently creating a methodology to validly distribute NetLogo, and ultimately Repast models to different nodes to create a series of small datasets that can then be merged into a single large dataset. In this way we hope to create a set of networks which can be combined in a meaningful way. Furthermore, as our understanding of the actual data improves, it is also clear that we must move the transaction network to a more scale-free structure with more discriminatory agents as our current network is vastly over connected.

3. Summary

Analysis of financial sector data presents a number of challenges. Some of these include its massive scale and dynamism. The MITRE Corporation is currently working on a set of tools to try an help the US Federal government overcome these challenges. As part of this work we are creating an agent based model that will create analogous datasets to the actual data being analyzed by the US Federal government. The reasons to create such a dataset include: possibility of better understanding the system that created the dataset, increased ability to collaborate with the community at large, test our relational graph analysis algorithms in a structured way, and give analysts a "sandbox" for what-if scenarios. Currently, the model creates datasets that contain a number of interesting structures and shows promise in creating the datasets necessary to test the graph analysis tools. As our understanding of the data increases it is becoming clear, however, that the model needs

to move to a more scale-free network structure that is much more sparsely connected that it is currently. These changes are in process. Finally, the infrastructure to run NetLogo on large scale cluster computers is in place. Now a methodology to distribute the network computation is being investigated so we may take full advantage of cluster computing as we attempt to create very large datasets.

4. References

[Axtell 2000] Axtell, R. (2000) "Why agents? On the varied motivations for agent computing in the social sciences", Center on Social and Economic Dynamics Working Paper No. 17, November 2000.

[Bloedorn 2005] Bloedorn, E., Rothleder, N., DeBarr, D., Rosen, L. (2005) "Relational Graph Analysis with Real-World Constraints: An Application in IRS Tax Fraud Detection." To appear in working notes from the AAAI-05 Workshop on Link Analysis. Pittsburgh, PA, July 2005.

[Collier & Howe 2003] Collier, N., T. Howe, et al. (2003) Onward and upward: the transition to Repast 2.0. Proceedings of the First Annual North American Association for Computational Social and Organizational Science Conference. 2003. K. Carley. Pittsburgh, PA, North American.

[Hazy & Tivnan 2004] Hazy, J. K. and Tivnan, B. F., (2004) "Simulation as Method in Organization Science." Proceedings of the Conference on Human and Organizational Studies, 257-262. Ashburn, VA: George Washington University.

[Koehler, et al. 2004] Koehler, M., Barry, P., Widdowson, B., Forsyth, A. (2004) "Case Study: Using Agents to Model Stability and Support Operations," Proceedings of Agent 2004. Chicago: Argonne National Laboratory.

[McKelvey 2002] McKelvey, B. (2002) "Model-Centered Organization Science Epistemology". Companion to Organizations, 752-780, J. A. C. Baum. Oxford, UK: Blackwell.

[Sterman 1992] Sterman, J. D. (1992). "Teaching Takes Off: Flight Simulators for Management Education," OR/ MS Today (October), 40-44.

[Wilensky 1999] Wilensky, U. (1999) NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL, 1999.

[U.S. PATRIOT ACT 2005] U.S. PATRIOT ACT HR 3162, § 602, http://www.fincen.gov/hr3162.pdf