Visualizing Multichannel Synthetic Data for Modeling and Simulation

Author: Patrick Corbett, Network Systems Science and Advanced Computing Division (NSSAC), Biocomplexity Institute, University of Virginia

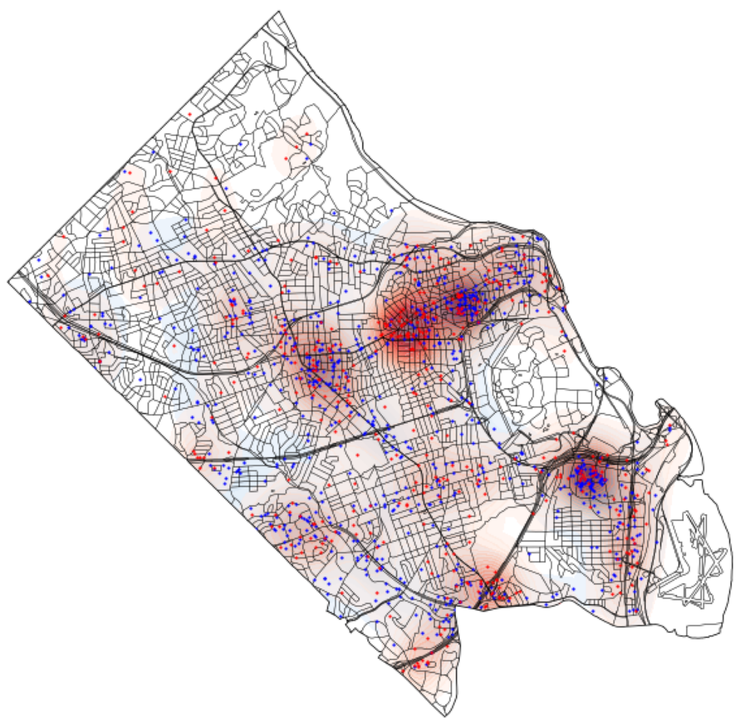
NSSAC Mentors: James Schlitt, Bryan Lewis, and Samarth Swarup

Background

Within this project we sought to model the behaviors and interactions of simulated individuals in Arlington, Virginia over a span of four weeks. To this end, we constructed a synthetic Arlington population using an integration of open and closed data sources including the 2010 census, police reports, household expenditure surveys, and time-use surveys.

- Simulated locations and times of activities were used to construct data channels including SMS messages, credit card payments, web searches, and police reports.
- Potential research applications include modelling energy consumption, the spread of infectious disease, the occurrence of chronic disease, and for identifying nefarious actors.
- To the best of our knowledge, this is one of the most comprehensive population interaction models in terms of scale, complexity, and temporal specificity.

Figure One: Day and Night Crime Reports

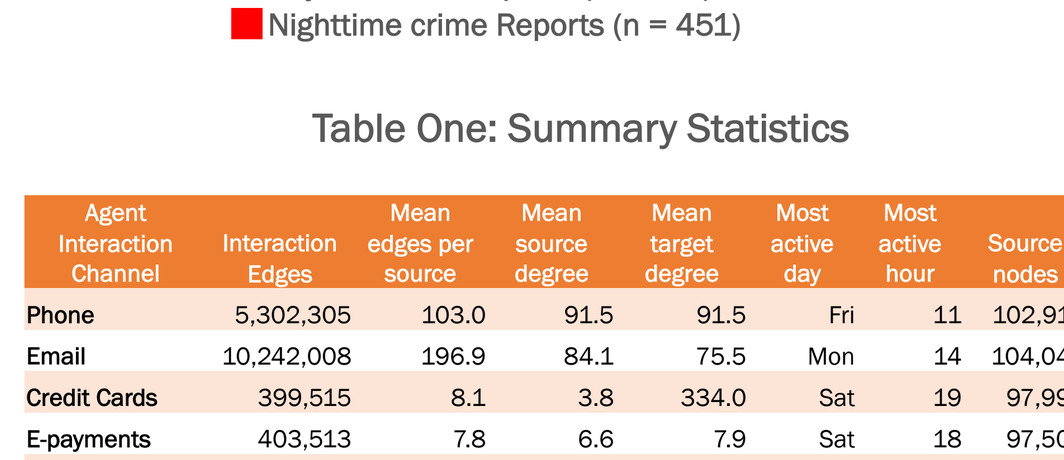


Daytime crime reports (n = 855)
Nighttime crime Reports (n = 451)

Table One: Summary Statistics

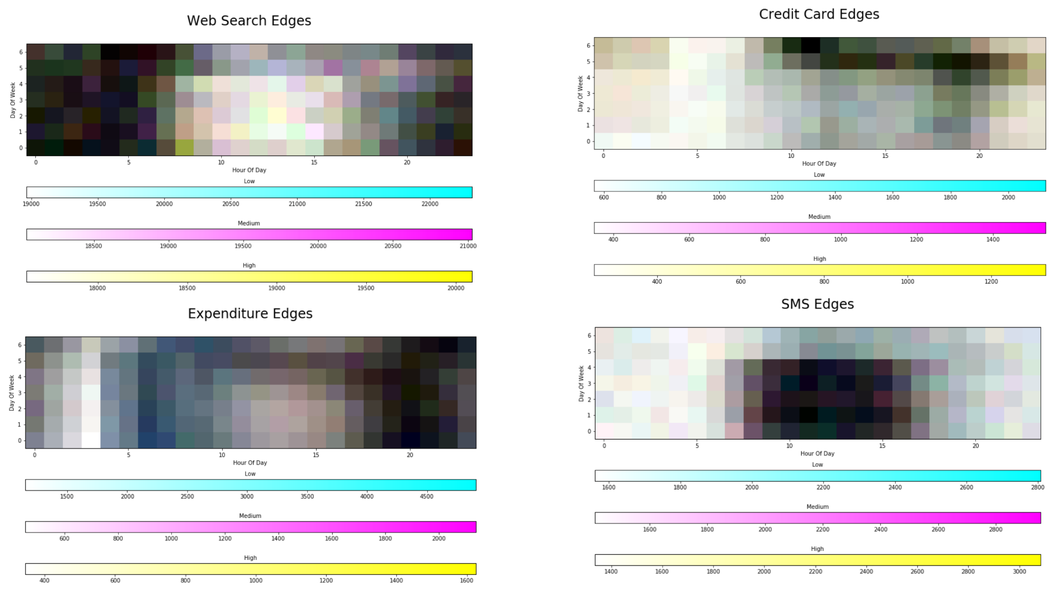
Agent Interaction Channel	Interaction Edges	Mean edges per source	Mean source degree	Mean target degree	Most active day	Most active hour	Source nodes
Phone	5,302,305	103.0	91.5	91.5	Fri	11	102,918
Email	10,242,008	196.9	84.1	75.5	Mon	14	104,046
Credit Cards	399,515	8.1	3.8	334.0	Sat	19	97,993
E-payments	403,513	7.8	6.6	7.9	Sat	18	97,504
SMS	18,508,265	359.7	303.5	303.5	Tue	11	102,918
Location	14,255,454	150.4	5.7	6.7	Fri	4	102,918
Expenditure	8,136,979	107.1	2.4	3.9	Sat	20	94,461
Web Searches	66,020,728	1,269.1	239.8	21,882.7	Sun	4	102,918
Police Reports (P-L)	1,306	1.2	1.0	1.1	Thu	15	1,179
Police Reports (C-L)	1,306	2.2	1.1	9.7	Thu	15	1,082
Venmo	7,024,852	2.0	1.9	1.9	Tue	22	4,537,546

Figure Two: Total Edge Counts



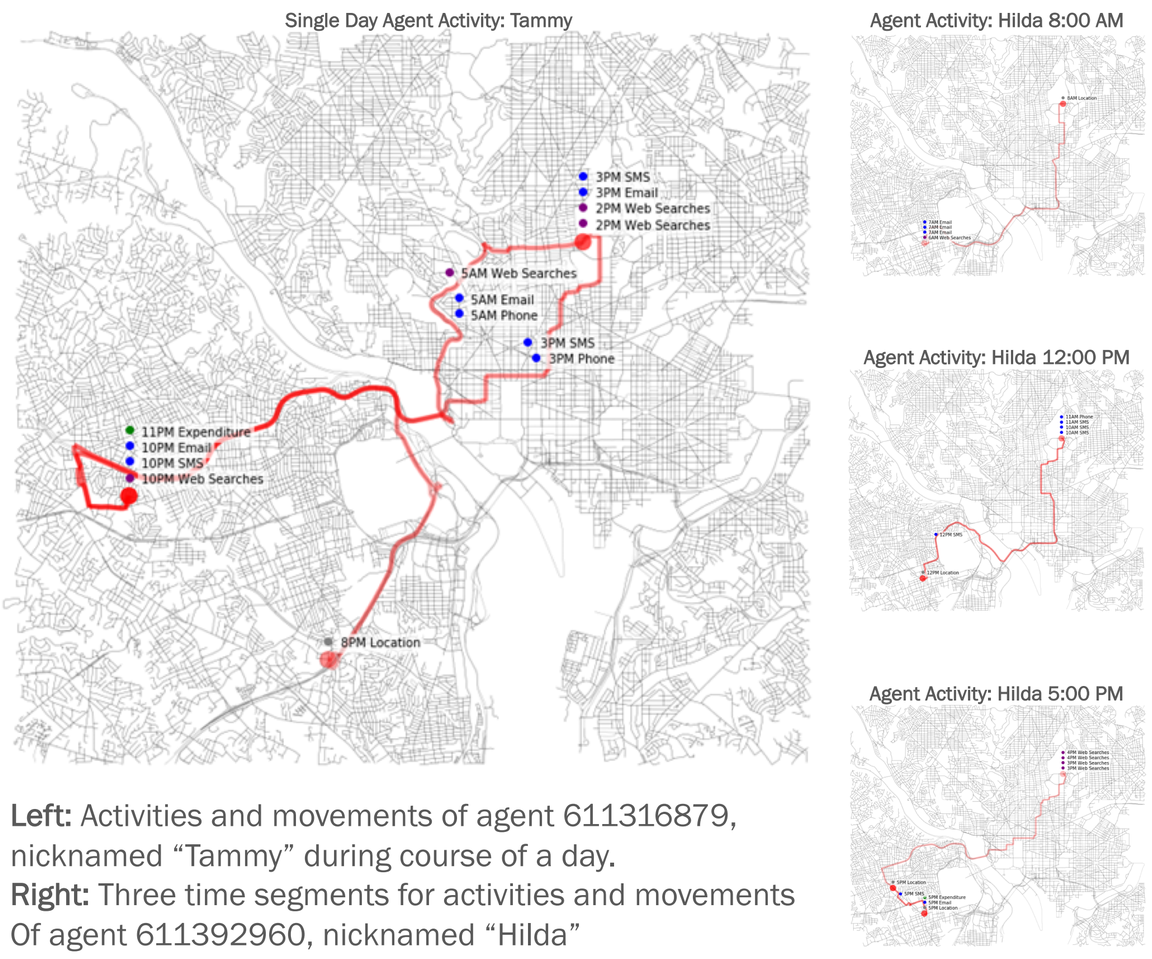
The total number of edges by the hour of day and day of the week for select channels.

Figure Three: Comparative Edge Counts



The total number of edges by the hour of day and day of the week for select channels sorted into tiers of low, medium, and high numbers of unique contacts.

Figure Four: Agent Stories



Left: Activities and movements of agent 611316879, nicknamed “Tammy” during course of a day.
Right: Three time segments for activities and movements Of agent 611392960, nicknamed “Hilda”

Conclusions And Future Work

One strength of this synthetic data is that it provides harmonized individual and population spatiotemporal data that is nearly impossible to obtain from the real world. As such, the data has broad implications with regard to its research applications within scientific modeling and simulation research.

- The above results indicate a dynamic virtual model that provides information regarding a vast number of individuals on a day to week timescale
- This model provides a framework for the development and evaluation of network algorithms intended to find patterns of suspicious or malicious behaviors of bad actors, such as drug dealers, amongst the copious digital signals our societies produce.
- Future work could include seeding bad actors into the simulated society and test detection methods based on unique combination of channel attributes.

References

- TIGER/Line Shapefile, 2013, county, Arlington County, VA. (2013, November 21). Retrieved August 4, 2019, from https://catalog.data.gov/dataset/tiger-line-shapefile-2013-county-arlington-county-va-current-topological- faces-polygons-with-al/resource/c9395320-acc1-44c7-92bf-5ea35ff03876?inner_span=True