UNIVERSITY OF VIRGINIA DATA SCIENCE INSTITUTE

MITRE

Socio-Temporal Trends in Urban Cultural Subpopulations through Social Media

The Problem

Background

- The United Nations predicts **6 BILLION** people will live in urban areas by 2045 [1]
- How can we better define an increasingly diverse urban fabric?
  - Understand *Where*, *How* and *When* people interact
  - Cities can best serve their citizens
    - Internationally and Domestically

State of the Art

- Individuals grouped into broad categories, like race
- Robust data collection is expensive and time-consuming
  - 10 years to compile the U.S. Census [2]
  - Inconsistent data sources inhibit universal methodologies
- Temporal, geographic or clustered trends
  - Rarely the intersection of these realms [3-5]
- Twitter's user base approximates Census data
  - Domestically and abroad [6-7]

Methodology

- Classify users more granularly than by race or income
- Analyze temporal interactions between groups
- Social Media offers a near-real-time pulse of a city
  - Readily available data streams worldwide
- Create a generalizable, region-agnostic approach
  - Test the methods in distinctly diverse cities
    - Los Angeles, Chicago, & Istanbul

I. Topic Modeling

Approach

- Combine each users' Tweets into a single document
- Aggregate users by their chosen Twitter language
- Latent Diriclet Allocation forms clusters based upon the frequency of co-occuring terms [8].

City	Raw Tweets	Filtered Tweets (Spambots removed)	Top User Languages
Los Angeles	10,476,477	9,063,602	English (98.1%), Spanish (0.8%)
Chicago	5,967,115	4,978,647	English (99.2%), Spanish (0.4%)
Istanbul	6,213,382	5,072,874	Turkish (89.3%), English (9.1%)

Table I - Summary of Twitter Data
Collected October 28, 2016 to January 28, 2017

Results

- 20 total topics per region, from languages accounting for at least 5% of all Tweets
  - 20 English topics in Los Angeles and Chicago
  - 17 Turkish & 3 English topics in Istanbul
- Languages not selected for Topic Modeling are treated as "Language-Topics"

City	Top 10 Words	Topic
Los Angeles	namm, songwriter, home, cali, music, producer, life, singer, soccer, night	"Music"
Los Angeles	la, hillary, obama, clinton, donald, voted, trump, american, russia, racist	"Politics"
Chicago	flythew, worldseries, gocubsgo, wrigley, united, insta, hamilton, field, canada, parade	"Cubs"
Chicago	indiana, basketball, Instagram, coach, varsity, football, central, county, official, baseball	"Sports"

Table II - Sample Topics
Terms appear in no more than 3.5% of all Documents in a Region

Conclusions

- Topic modeling is generalizable
  - Works in multiple locations
  - Support for multiple languages
- Language/Topical classifications
  - More granular than demographics
- Low thresholds produce better topics
- Potential Improvements:
  - Vectorized word representation
  - Algorithms for sparse text

II. Time Series

Approach

- Produce ARIMA Models for the topics generated above
  - Auto-regressive Integrated Moving Average
- Heirarchical Clustering of Top 20 Topics
  - Similarity metric - correlation between topics
  - Complete link - maximize distance between groups

Conclusions

- Heirarchical Clustering aggregates similar-moving topics
  - Arabic (Green) is its own cluster in LA and Istanbul
  - Bilingual topics in Chicago and Los Angeles
    - English topics containing Spanish stop words
    - Separate from all other Spanish and English topics
- National and Regional Events influence Twitter volume
  - Increase: Game 7 of the World Series, U.S. Election
  - Decrease: Turkey restricts internet access [9].
- 61.7% of ARIMA models lack stationary
  - Fluctuating variance or a linear trendv

Results

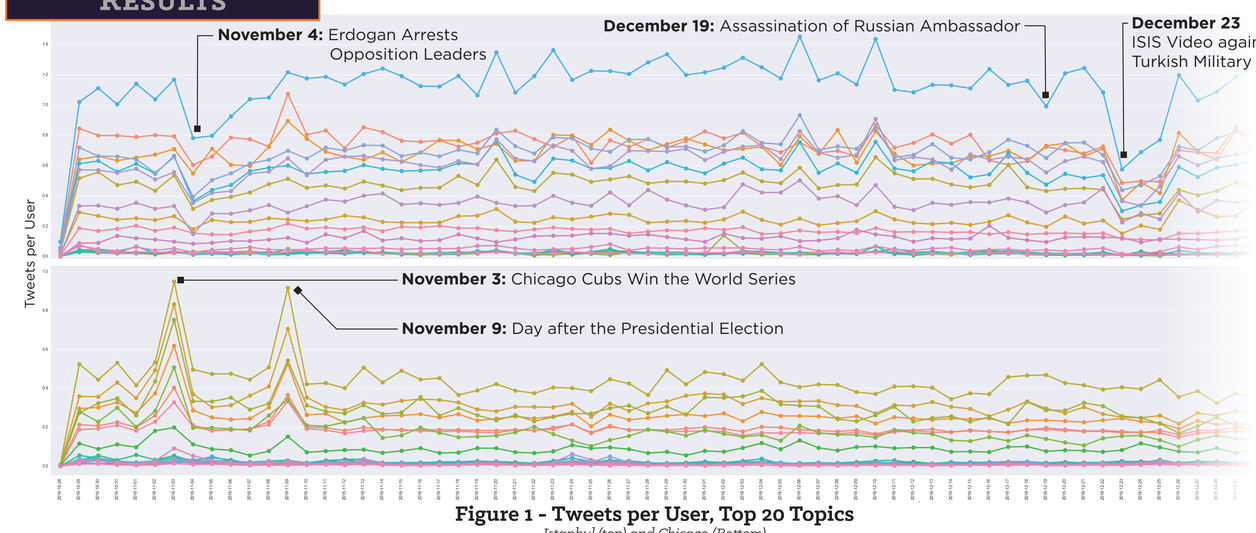

Figure 1 - Tweets per User, Top 20 Topics
Istanbul (top) and Chicago (Bottom)

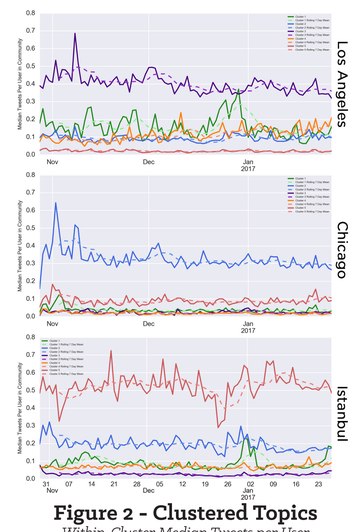

Figure 2 - Clustered Topics
Within-Cluster Median Tweets per User

III. Louvain Clustering

Approach

- Identifies latent clusters in large networks [10]
- Each node is an individual user
  - Connected by weighted cosine similarities of term frequency (TF*IDF) matrix
- Highly interconnected nodes/users form communities
- Optimized Louvain networks feature:
  - Dense connections within the same communities
  - Sparse links to users in other communities

Conclusions

- Connections and interactions between groups emerge.
  - Subcommunities exist within each group.
- Native language: Tightly interconnected communities
- Foreign languages: Distinct Communities

Results

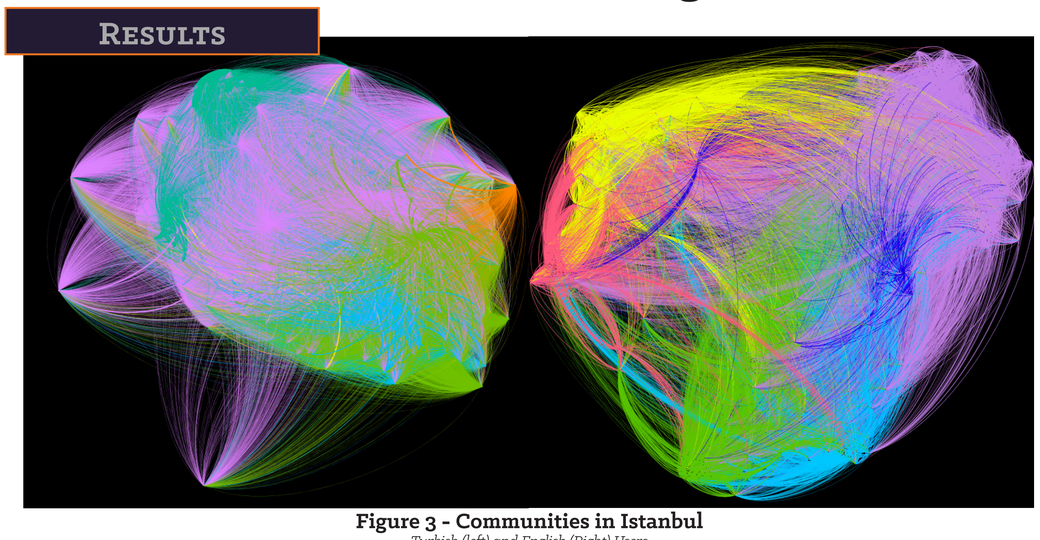

Figure 3 - Communities in Istanbul
Turkish (left) and English (Right) Users

Language	Color	Sample Words
Turkish	Blue	socialism, beşiktaş(football club), mosque, iman, kadıköy (neighborhood), away, sport fans, wife, will not happen, your future, galatasarai (football club)
Turkish	Purple	life, meaning, arda (Arda Turan, Erdogan supporter), you, own, mother, even, fair, atatürkçü (secular organization that espouses the ideas of the founder of modern Turkey), Republic, woman
English	Yellow	emi (European Movement International), universal, music, Turkey, office, formal, account, old, group, pick, January, during, here
English	Purple	core, bad, from, people, displeasure, myself, kemalism (nationalist ideology), Turkish, world, race, expression, make, purpose, using

Table III - Sample Louvain Communities
Terms appear in no more than 3.5% of all Documents in a Region

IV. Geographic

Approach

- Geotagged Tweets parallel Census findings [6-7, 11].
- Kernel Density Estimates (KDE) locate groups in space

Conclusions

- Preferred Twitter Client varies by region
  - Instagram generalizes location, creating artificially saturated points.
- Events (i.e. New Year's Eve) have a profound effect
- Groups have unique denisity patterns
  - Densities ebb and flow through time and space

Results

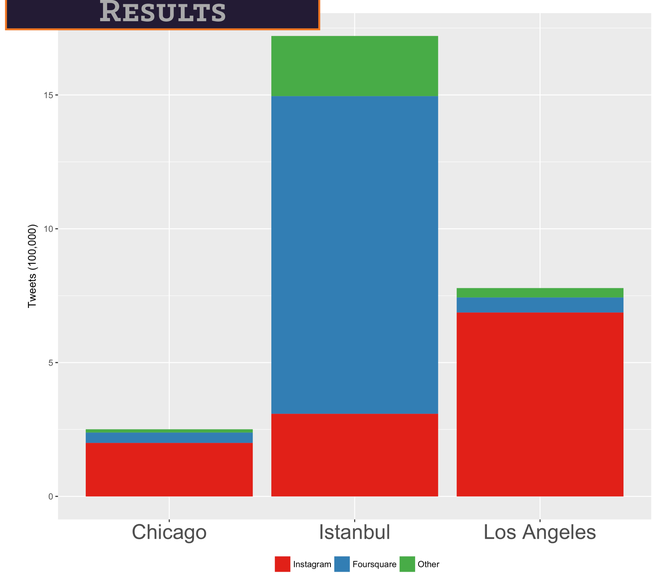

Figure 4 - Sources of Geolocated Tweets
Istanbul prefers Foursquare, Chicago and Los Angeles use Instagram

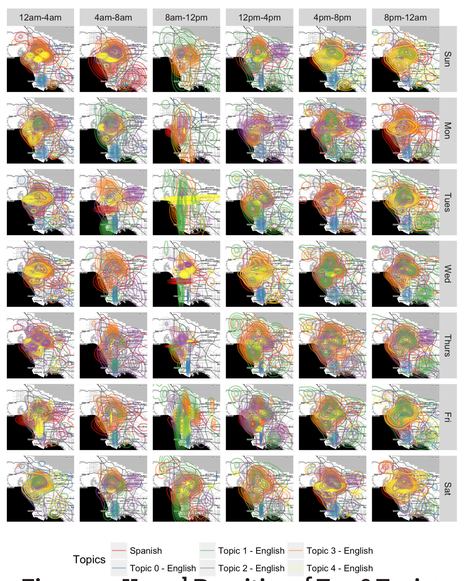

Figure 5 - Kernel Densities of Top 6 Topics
Los Angeles by Day (rows) by Hour (columns)

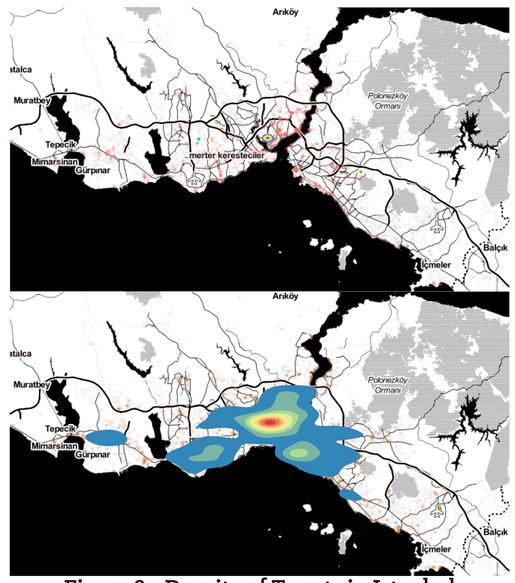

Figure 6 - Density of Tweets in Istanbul
New Year's Eve (top) and all other Sundays (bottom), 12am-4am

Authorship

Lander Basterra (llb2eu@virginia.edu)
Tyler Worthington (tjw4ry@virginia.edu)
James Rogol (rogol@virginia.edu)
Data Science Institute, University of Virginia

Special Thanks to:

The Mitre Corporation, Dr. Donald Brown, Mohammad Al Boni, Ahmed Asker, and the Faculty & Cohort at the Data Science Institute

References

[1] United Nations Department of Economic and Social Affairs. July 2014. "2014 Revision of World Urbanization Prospects." eas.un.org/unpd/wup/.

[2] United States Census Bureau. 2000. "Factfinder for the Nation."

[3] Cable, Dustin A.; Martin-Anderson, Brandon; and Fisher, Eric. "The Racial Dot Map: One Dot per Person." July 2013. demographics.coopercenter.org/Racial-Dot-Map/.

[4] McKenzie, Grant; Janowicz, Krysztof; Gao, Song; et al. 2015. "POI Pulse: A Multi-Granular, Semantic Signature-Based Approach for the Interactive Visualization of Data." Cartographica 50(2), pp. 71-85.

[5] De la Rosa, Kelvin; Shah, Rushin; Lin, Bo; et al. "Topical Clustering of Tweets." July 2011. Proceedings of the ACM Special Interest Group on Information Retrieval 3rd Workshop on Social Web Search and Mining, Beijing, China.

[6] Stiger, Enrico; Westerholt, René; Resch, Bernd; and Zipf, Alexander. September 2015. "Twitter as an Indicator for the Whereabouts of People? Correlating Twitter with UK Census Data." Computers, Environment and Urban Systems 54, pp. 255-256

[7] Steiger, Enrico; Ellersiek, Timothy; and Zipf, Alexander. November 2014. "Explorative Public Transport Flow Analysis from Uncertain Social Media Data." Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, Dallas, TX, pp. 1-7.

[8] Blei, David M.; Ng, Andrew Y.; and Jordan, Michael I. March 1, 2003. "Latent Dirichlet Allocation." Journal of Machine Learning Research 3, pp 993-1022.

[9] Turkey Blocks. 2017. "Mapping Internet Freedom in Real Time." turkeyblocks.org.

[10] Blondel, Vincent; Guillaume, Jean-Loup; Lambiotte, Renaud; and Lefebvre, Etienne. 2008. "Fast Unfolding of Communities in Large Networks," Journal of Statistical Mechanics: Theory and Experiment, pp. 2-7.

[11] Gao, Song; Yang, Jiue-An; Yan, Bo; et al. September 2014. "Detecting Origin-Destination Mobility Flows from Geotagged Tweets in the Greater Los Angeles Area." Proceedings of the 8th International Conference on Geographic Information Science, Vienna, Austria.