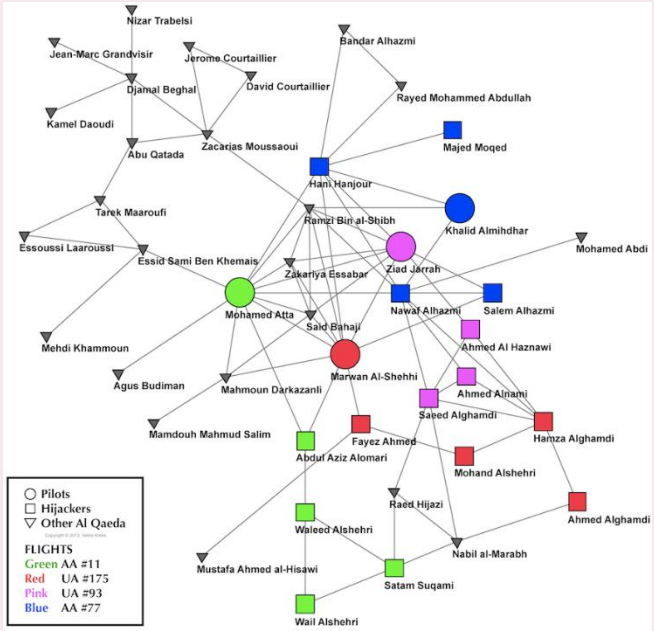
Call Data Records

Lecture 28: CSE 490c



Topics

- Data Science for Development
- AI for Social Good

- Today
  - Call Data Records

Announcements

- Homework 7 Due Tonight
- Programming Assignment 4 Due December 11
- You have received the teaching evaluation link. Complete evaluations by December 9

A high response rate is very important for meaningful results. We will send reminder emails to non-responders during the evaluation period. In addition, studies show that instructor involvement can increase response rates as much as 15-20%.

Telco Information

- Call Data Records (Call Detail Records)
  - Meta data on individual calls
- Cell Tower Logs
  - Information of handset connections with towers
- Cell Tower Locations

Access to cell phone data

- Proprietary to Telco
- Provide competitive advantage
  - Possibly for marketing or data services
  - Linkage with mobile money
- Subject to government privacy regulations
- Possible access to aggregated data

Reading for this week

- An Investigation of Phone Upgrades in Remote Community Cellular Networks, Kushal Shah et al., ICTD 2017

Call Data Records

- Meta data associated with calls
- Source number
- Destination number
- Source Tower (ID)
- Destination Tower (ID) [might be missing]
- Time
- Duration
- Status

Call Detail Records

LATITUDE	LONGITUDE	DATE	TIME	NUMBER	NAME	DURATION
44.50880 N	73.18223 W	1/28/2008	0917	802-555-1234	Chittenden Bank	0:10:17
44.50880 N	73.18223 W	1/28/2008	0942	802-555-8673	Poopsie LaRue	0:01:03
44.50880 N	73.18223 W	1/28/2008	0945	802-555-9201	Hanley Strappman	0:05:32
44.27834 N	73.21263 W	1/29/2008	2205	802-555-7758	Verizon Voice Mail	0:01.13
44.27834 N	73.21263 W	1/29/2008	1532	802-555-4492	Widgets LLC	0:03:47
44.27834 N	73.21263 W	1/29/2008	2209	802-555-7758	Verizon Voice Mail	0:00.36
44.50880 N	73.18223 W	1/30/2008	0830	202-555-1818	British Embassy	0:18:12
44.27834 N	73.21263 W	1/30/2008	2208	802-555-7758	Verizon Voice Mail	0:00.53
44.27834 N	73.21263 W	1/30/2008	2211	802-555-8673	Poopsie LaRue	0:06:18
44.50880 N	73.18223 W	1/31/2008	0903	202-555-1843	British Embassy	0:03:21
44.50880 N	73.18223 W	1/31/2008	0908	416-555-9834	British Embassy	0:22:04
44.4143 N	73.03561 W	1/31/2008	1047	802-555-9201	Hanley Strappman	0:01:02
44.4143 N	73.03561 W	1/31/2008	1050	213-555-2761	M. Fendell	0:09:06
44.25295 N	72.58229 W	1/31/2008	1127	802-555-9201	Hanley Strappman	0:05:38

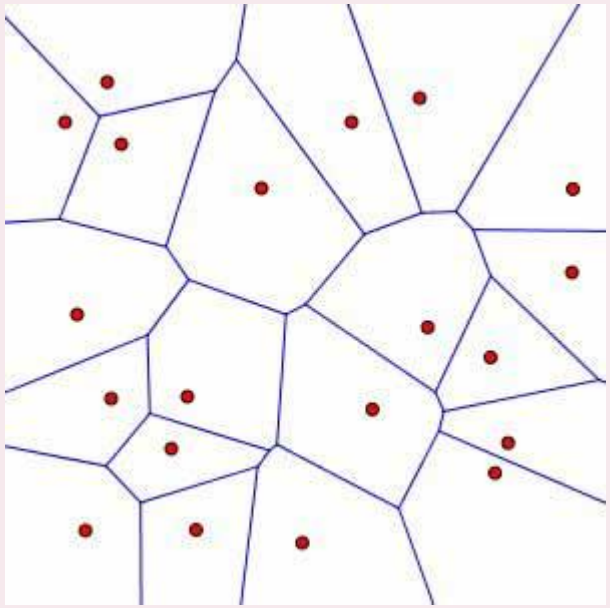
Types of studies

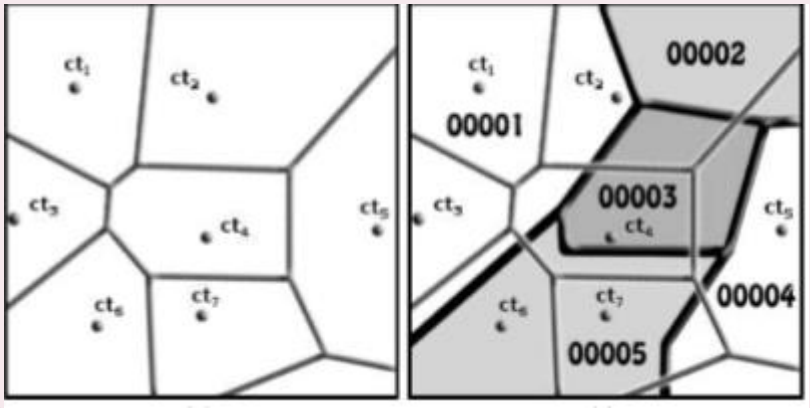
- How people use technology
- Populations studies (where people are)
- Event studies (what happens when)
- Epidemiology studies
- Economic studies

Distinction between Aggregate Studies and deriving information about individuals

Working with CDRs

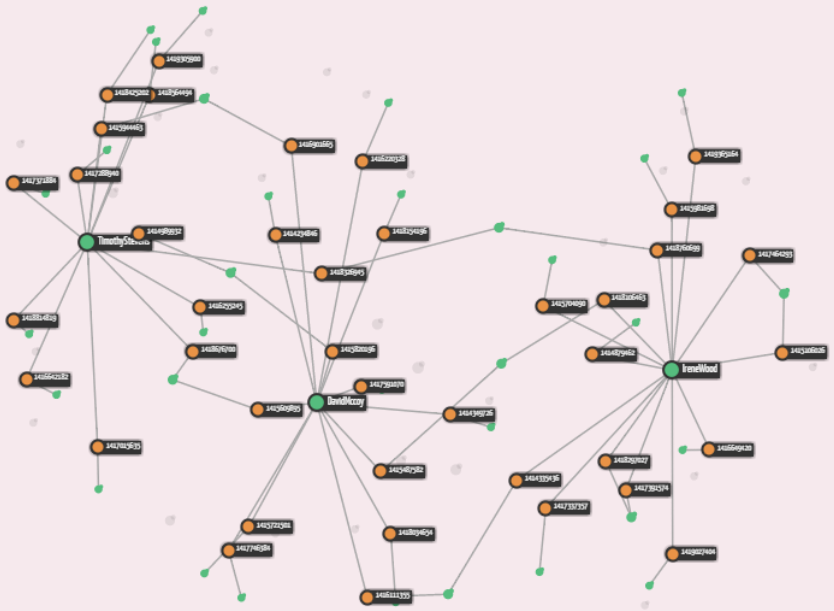
- Preprocess data for higher level structure
- Align data with other sources
  - Tower data
  - Economic / Population Data
- Compute home location
- Determine movement patterns





Call Graph

- Directed or undirected graph on calls
- Measurement of call volume
- Detection of high in-degree and out-degree nodes
- Identification of social network

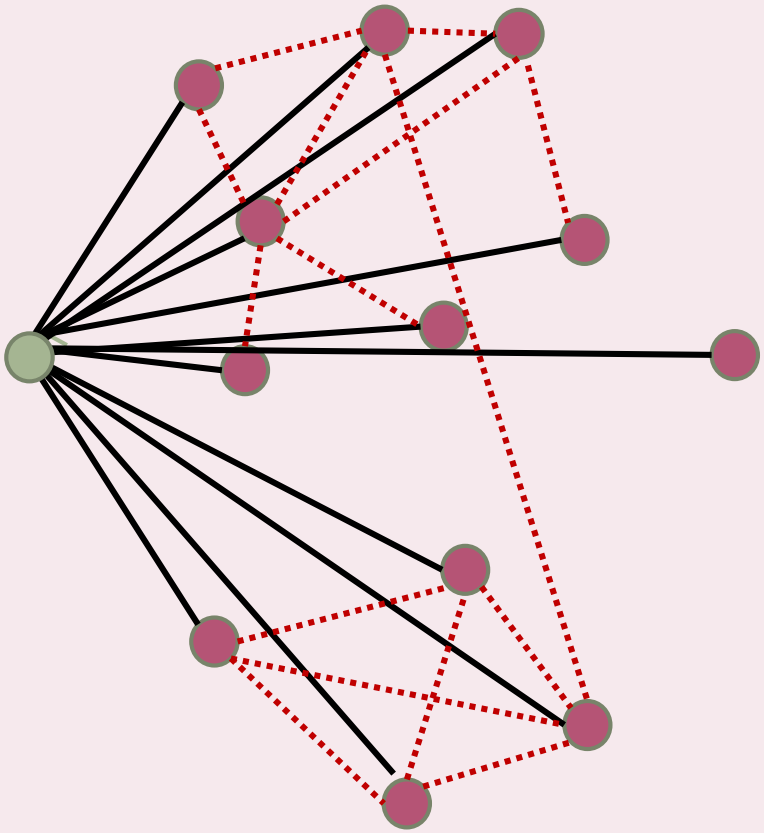


Call Graph Analysis

- Global analysis versus individual analysis
  - Is the goal to understand aggregate properties or individual properties
- Feature identification of individual's calls
  - Number of calls
  - Length of calls
  - Missed calls
  - Incoming vs out going
  - Time of day
  - Neighborhood
    - Frequent caller neighborhood
  - Nodes of distance two

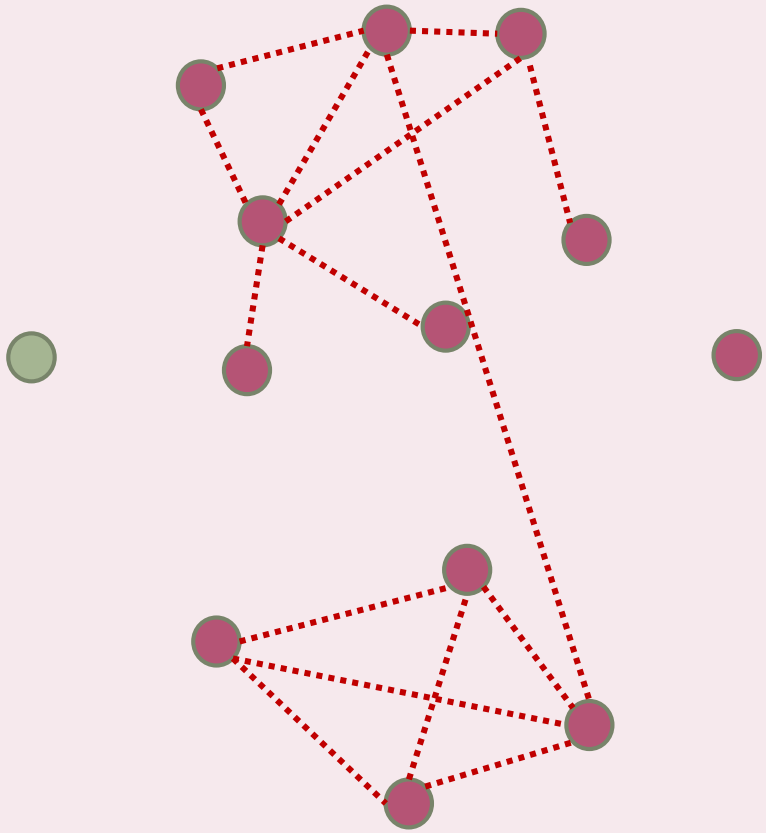
Social Network Identification I

- How do you identify a callers Social Network from a call graph?
- Start with the Induced Subgraph on the Neighborhood of the individual



Social Network Identification II

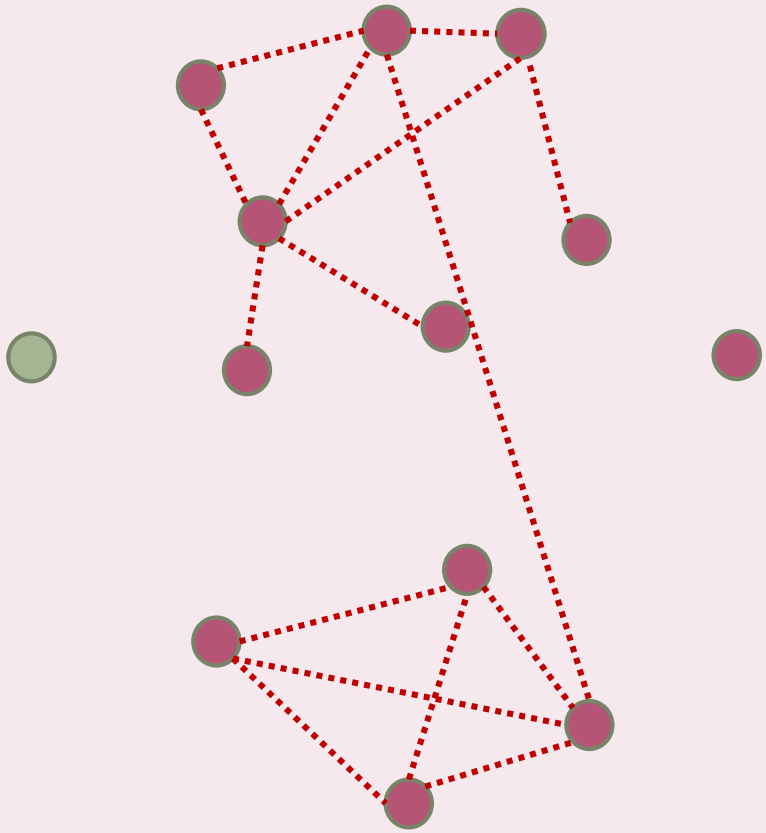
- Identify highly connected groups of vertices in Neighborhood Graph
- Finding a maximum Clique is NP-Complete
- Heuristics
  - Maximum degree subgraph
  - Maximum density subgraph



Degree-K Subgraph problem

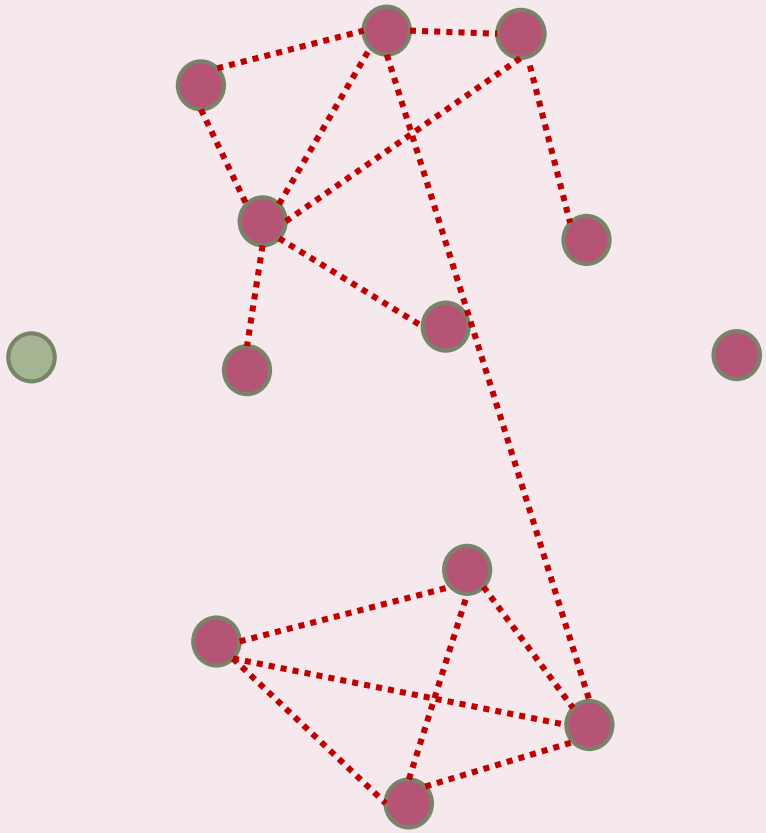
While there is a vertex of degree less than K

- Delete all vertices of degree less than K



Maximum Density Subgraph problem

- Find an induced subgraph S that maximizes ration Edge(S)/Vertices(S)
- Polynomial time algorithms using Network Flow techniques
- Related to degree K subgraph problem



How good a predictor is the call graph of an individual's income?

Phone use studies

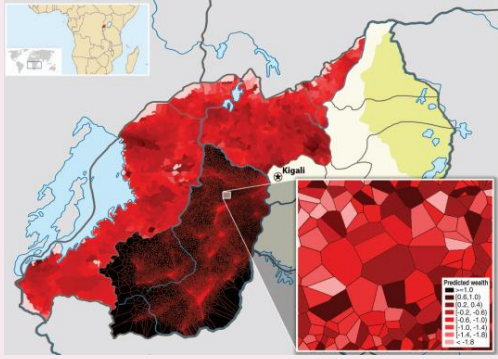
- Given demographic information, study phone use behavior
  - Call volume, call timings (time of day, date), number of contacts
- Studies of Sim Card Churn
- Inference of demographics from behavior

Mehrotra et al. (2012), Differences in Phone Use Between Men and Women: Quantitative Evidence from Rwanda.

Migration Studies

- Movement of people is an area of significant study
- Lack of census data make this hard to study
- Short term migration
  - What are the patterns
  - Is it possible to distinguish between shorter term and permanent migration
- Forced migration and droughts
  - Match to climate data
  - Question on local versus long distance migration
- Technical issues in definitions of movement

Blumenstock et al. (2015), Predicting poverty and wealth from mobile phone metadata



Economic Studies

- Predict economic status at a local (e.g. District) level
- Household surveys are expensive. Idea is to use Cell Phone Data to expand surveys
- Correlate household surveys with CDR
  - Compute wide range of properties of CDR
  - Construct machine learning model to predict household assets (from survey data)
  - Apply to all call records in the data set

Epidemiology

- Correlating human movement data and disease frequency
- Substantial work on Malaria and Call Data Records
- Key use case is malaria elimination
  - Understanding if cases are local infections or from other regions
  - Understand movement from high incidence to low incidence areas
- Technical modelling work that combines migration and economic studies

Event studies

- Look at impact of events in data sets
- High volume of calls related to disasters, elections, holidays
- Spike in call volumes has been observed associated with earth quakes
- Significant interest in call data records and disaster response
  - Technical issues related to infrastructure and economic displacement

Additional challenges on CDR Analysis

- Countries often have multiple Telcos
  - Getting data from all Telcos is even harder
  - Is data from one Telco sufficient?
- Increasing use of other media for communication
  - Encrypted messaging Apps such as WhatsApp
  - Analysis of Social Media company data should be even more restricted than CDR
    - Aleksandr Kogan