

Fast Class Partitioning for Large Social Networks

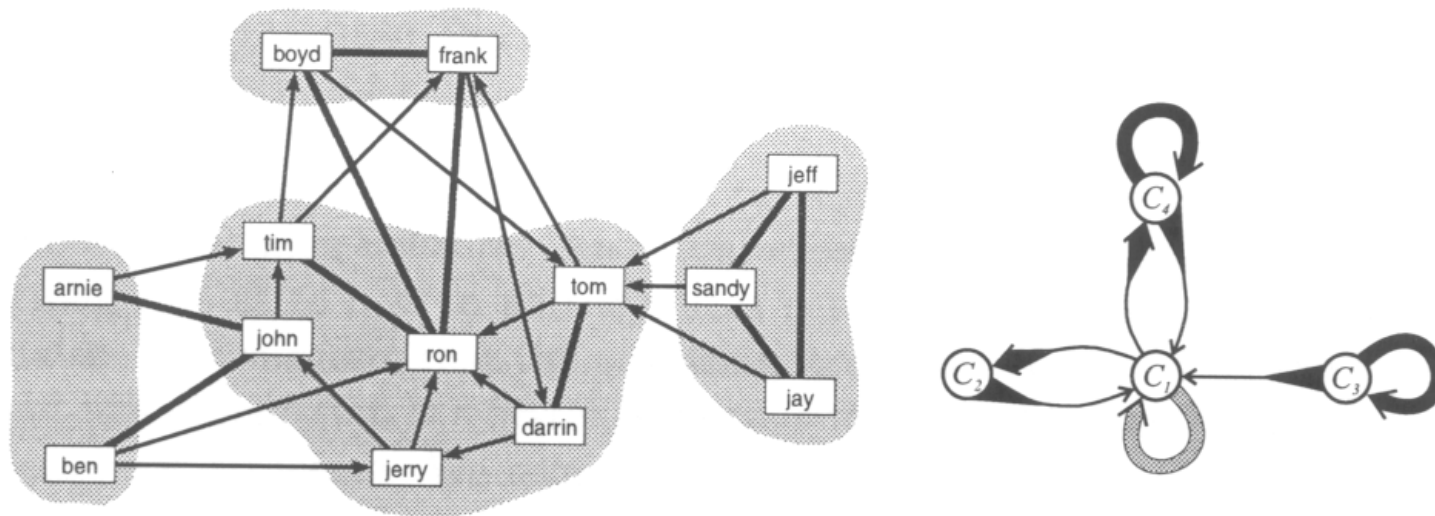
Salvatore J. Babones
Department of Sociology
University of Pittsburgh
sbabones@pitt.edu

INTRODUCTION TO SOCIAL NETWORK ANALYSIS

THE PURPOSE OF SOCIAL NETWORK ANALYSIS IS TO PARTITION THE INDIVIDUALS IN A NETWORK INTO STRUCTURALLY EQUIVALENT CLASSES

- Social network analysis grew out of the sociometric study of small groups
 - Ties between individuals are represented as edges (arcs) in a graph (digraph)
 - Typical study groups are school classes, summer camps, scientific collaborators
- Each member of the group is asked to identify his/her ties to others in the group
 - Typical questions are: “Whom do you know personally?”, “Who are your best friends?”, “Whom do you most respect?”, “With whom would you most like to be friends?”
 - Ties are usually coded 0/1, but can be given magnitudes
- Clustering techniques are then used to group structurally similar individuals
 - Originally this was done algebraically by repeatedly swapping rows and columns
 - A newer, faster technique is CLHC with ties to each individual taken as a dimension

A TYPICAL SOCIOMETRIC EXAMPLE IS A NETWORK ANALYSIS OF A LITTLE LEAGUE BASEBALL TEAM . . .



- Taken from Patrick Doreian, Vladimir Batagelj, and Anuska Ferligoj. 2003. "Positional Analyses of Sociometric Data." Working paper, Univ. of Pittsburgh.

SUCH SOCIOMETRIC TECHNIQUES ARE NOT VERY USEFUL FOR THE STUDY OF LARGE, OPEN NETWORKS

- It is only practical to study small ($n \ll 1000$) networks
 - The most effective clustering algorithms are worse than exponential in the number of cases
 - Visualization of data/results is challenging for large networks
- Full information is assumed for all ties among all subjects
 - Each subject must voluntarily reveal his/her social ties
 - Ties between individuals are given an inappropriate level of concreteness
 - All social ties outside the (closed) study group are ignored
- There is no sampling theory available, or really even conceivable

FRAMING THE PROBLEM

DATA ON REAL-WORLD SOCIAL NETWORKS OF INTEREST ARE LIKELY TO ARISE FROM TIME-SAMPLED OBSERVATIONS OF GROUP INTERACTION

- **HACKER COMMUNITIES:**

- Chat rooms can be monitored at fixed or random intervals and participants in the room logged
- Common blocks of code can be used to tie hackers to each other
- Hits from multiple sources to the same address can be linked together

- **TERRORIST NETWORKS:**

- Telephone monitoring to establish the existence of a conversation uses far fewer resources than actually listening to the conversation
- Automated e-mail monitoring can log the list of addresses mentioned in the same e-mail

- In each case the resulting data are a series of “snap-shots” of who is connected to whom at a given time or place

IMPORTANT NETWORK STRUCTURES MUST BE TEASED OUT FROM “DATA OF CONVENIENCE”

- Network partitioning algorithms should be robust with respect to sampling
 - At a minimum, there should be generalizability from random time-sampling of a network
 - Ideally, partitioning should be robust even when portions of the network are systematically over-sampled
- Another key objective should be to distinguish true network participants from “innocent bystanders”
 - Clustering algorithms should tie together individuals who interact among themselves
 - Unitary side-branches (e.g., “the terrorist’s mother”) should be ignored
- Emphasis should be placed on the study of “peer-to-peer” networks
 - Hub-and-spoke networks are easy to investigate using traditional means
 - Centrality in webs is a much more difficult concept to measure

ANALYTICAL TECHNIQUES SHOULD BE RULE-BASED TO MINIMIZE THE NEED FOR EXPERT INVOLVEMENT

- The most interesting and important datasets for open network analysis are likely to be very large
 - Relationships among subjects will be sparse
 - Networks will be partitioned into very large numbers of loosely-connected groups
- Rule-based techniques should be used for data preparation, clustering, and interpretation
- For a given known “individual of interest,” we would like to be able to look up:
 - The list of all those who are members of a similar structural partition in the network
 - The identities of the likely leaders within that partition
 - A list of observed subjects that are most likely to be aliases for the individual of interest

A TOP-DOWN APPROACH

INSTEAD OF WORKING UP FROM THE LEVEL OF THE INDIVIDUAL TIE, I PROPOSE TO WORK DOWN FROM THE LEVEL OF SOCIETY

- A SOCIETY (community of individuals) comprises the entire target population of the study
 - The goal is to partition a society into structurally equivalent CLASSES through iterative partitioning of existing classes
 - Each class represents a relatively closed “group of groups” within society
- In early partitions, we would like to hive off individual social ISOLATES:
 - Those who are found in groups with only with one other member of the community
 - Those who interact uniformly at random across all members of the community
- In later partitions, we would like to hive off aggregate social isolates

DATA MATRICES FOR SUCH GROUP-BASED NETWORK ANALYSES ARE RECTANGULAR RATHER THAN SQUARE

- The data are an $N \times I$ matrix of zeros and ones:
 - N is the number of individuals who compose the society of interest
 - I is the number of groups that have been observed
 - EACH NEW OBSERVATION REPRESENTS A NEW GROUP, EVEN IF IT CONTAINS THE SAME INDIVIDUALS AS A PRIOR GROUP!
- The data matrix can be used raw or normed
 - Norming is appropriate to adjust for differential group sizes and individual activity rates
 - I suggest dividing each entry by the column sum (the number of members in the group)
- Clustering is based on the individuals as cases and the groups as variables

PARTITIONING SHOULD BE BASED ON A SOCIAL DISTANCE MEASURE

- In traditional network analysis, each subject is coded for ties to each other subject
- I propose instead that commonalities between individuals be represented by the frequency of overlapping group memberships
 - This is more in keeping with conventional clustering analysis / profiling
 - This fits the data: at any time, groups of interacting individuals are observed together
- Each group membership is thus a dimension for clustering purposes
- A good measure of social proximity is the cosine between individuals' vectors
 - For those with no overlapping group memberships, the cosine is zero
 - Social distance is the opposite of proximity, operationalized by $1 - \text{cosine}$

GOOD PLACES TO STOP PARTITIONING ARE WHEN THE CLASS BEING HIVED OFF AS AN “ISOLATE” IS SIMILAR IN SIZE TO ITS SIBLING

- At each partitioning, a distant class is hived off the main body of the community
 - Complete linkage nearest neighbor clustering takes account of data vector values found throughout the entire community
 - Using the cosine as a proximity function eliminates the problem of scaling by degree of activity
- Each partitioning hives off a class of individuals who share extreme values from the perspective of the rest of society
 - As long as such a group is “small,” it is meaningful to label it an “isolate”
 - When a partitioning divides a class into two equal halves, it is not meaningful to label either an “isolate”
- Multiple stopping points may be substantively meaningful

THIS ANALYTICAL STRATEGY IMPROVES ON TRADITIONAL SOCIAL NETWORK ANALYSIS FOR STUDYING LARGE, OPEN NETWORKS

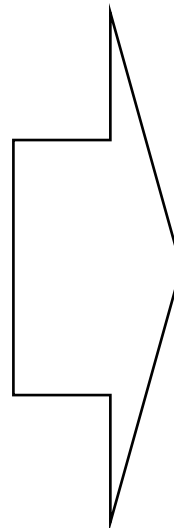
- It is practical to study large networks in an automated fashion
 - Clustering algorithms are typically polynomial in the number of subjects
 - Partitioning of subjects by class membership closely approximates common-language meanings of “networks” within a “community”
- Very little information is required about subjects
 - The strategy accommodates a data mining environment in which subjects need not know they are even being studied
 - There is no need for artificial constructs to create edges between individuals
 - Out-of-community individuals caught in the data are easily weeded out in analysis
- Amenability to sampling is obvious!

APPLIED ILLUSTRATION

I ATTEMPTED TO EXTRACT SECTION CLUSTERS FROM LISTS OF PARTICIPANTS AT SESSIONS AT THE JSM IN SAN FRANCISCO

- The American Statistical Association provided an electronic copy of the listing at the back of the 2003 program:

Abbott,Owen	453					
Abebe,Asheber	54					
Abeywickrama,KamalH.	376					
Abi-Habib,Natalie	443					
Abraham,KatharineG.	153	343				
Abramson,FlorenceH.	46					
Adak,Sudeshna	18	96	451			
Adams,AlyceS.	167					
Adams,Ann	409					
Adams,JohnL.	403					
Adams,TamaraS.	198					
Adlakha,Arjun	348					
Adler,Gail	465					
Adler,MicheleC.	16					
Afshartous,David	444					
Agans,Robert	41					
Agarwal,DeepakK.	293					
Agarwal,Sameer	3					
Agrawal,NancyG.	296					
Agrawal,Rakesh	156					
Agresti,Alan	104					
Agustin,MaZenian.	422					
Agustin,MarcusA.	169					
Ahn,Chaehyung	402					
Ahn,ChulH.	212					
Ahn,Hongshik	423					
Ahsanullah,Mohammad	126					
Aikens,Lynn	98	134	219	255	306	379
Ajmani,Vivek	238					
Akritis,MichaelG.	209	418				



SUBJECT	g1	g2	g3	g4	g5
Abbott,Owen	0	0	0	0	0
Abebe,Asheber	0	0	0	0	0
Abeywickrama,KamalH.	0	0	0	0	0
Abi-Habib,Natalie	0	0	0	0	0
Abraham,KatharineG.	0	0	0	0	0
Abramson,FlorenceH.	0	0	0	0	0
Adak,Sudeshna	0	0	0	0	0
Adams,AlyceS.	0	0	0	0	0
Adams,Ann	0	0	0	0	0
Adams,JohnL.	0	0	0	0	0
Adams,TamaraS.	0	0	0	0	0
Adlakha,Arjun	0	0	0	0	0
Adler,Gail	0	0	0	0	0
Adler,MicheleC.	0	0	0	0	0
Afshartous,David	0	0	0	0	0
Agans,Robert	0	0	0	0	0
Agarwal,DeepakK.	0	0	0	0	0
Agarwal,Sameer	0	0	1	0	0
Agrawal,NancyG.	0	0	0	0	0
Agrawal,Rakesh	0	0	0	0	0
Agresti,Alan	0	0	0	0	0
Agustin,MaZenian.	0	0	0	0	0
Agustin,MarcusA.	0	0	0	0	0
Ahn,Chaehyung	0	0	0	0	0
Ahn,ChulH.	0	0	0	0	0
Ahn,Hongshik	0	0	0	0	0
Ahsanullah,Mohammad	0	0	0	0	0
Aikens,Lynn	0	0	0	0	0
Ajmani,Vivek	0	0	0	0	0
Akritis,MichaelG.	0	0	0	0	0

- I rearranged this data into a rectangular matrix of individuals observed in groups

USING BOTH RAW AND NORMED MATRICES, I PARTITIONED THE JSM COMMUNITY INTO (SADLY NOT VERY INTERESTING) CLASSES

- I used complete linkage nearest neighbor to cluster, with cosine as the distance
- Raw and normed matrices yielded very similar partitions
- Unfortunately, the resulting partitions corresponded to one for each “observed” session at the JSM
- Further analysis revealed that there is virtually zero overlap among session participants at JSM sessions - in essence, there were no “usual suspects”
- On the bright side . . . the negative results were a correct characterization of the class structure of the American Statistical Association as reflected at the JSM

TWO CONJECTURES

CONJECTURE: CLASS LEADERS CAN BE IDENTIFIED ENDOGENOUSLY

- A good measure of class leadership is probably:

Leadership index = [Norm of an individual's data vector] *

[Cosine of the angle between the individual and the class center]

- This measure can be computed for all individuals regardless of class membership
 - The structure of the data implies that classes must separate by dimension, rather than by degree
 - Thus, non-members of a class will have low or zero cosines
- Very active individuals who participate in groups across the spectrum of classes will receive low leadership scores on all classes

CONJECTURE: CLASS PARTITIONINGS SHOULD BE ROBUST WITH RESPECT TO A CONSTANT RATE OF IDENTITY SHIFTING

- Classes are identified by overlapping memberships among multiple groups
 - In other words, we see “the usual suspects” together over and over again
 - These “usual suspects” cluster into a class, while “bystanders” fall off at right angles
- When users take aliases, they seem to be new bystanders
- However, new groups continue to be observed that contain many of the usual suspects under their original identities, giving continuity to the class
- The resulting classes (with identity shifting) will seem larger in membership
 - On average, each member will seem to be less engaged in the class
 - Thus, the norm of the class center will be smaller in magnitude, but identical in direction