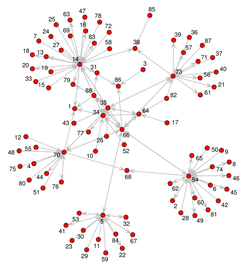
Dynamic Graphs in Communication Networks

Chipman, H. and Nettel-Aguirre, A.

Mathematics and Statistics Department, Acadia University, Wolfville, NS

Introduction

- Communications in a network are usually interpreted as a graph with nodes and edges.
  - Nodes represent members,
  - edges represent communications between members.
- Graphs (and networks) are considered dynamic as they change through time.
- Groups of nodes in the graphs form communities of interest (COI). An example is seen in the figure below



- Communities of interest can change:
  - The pattern of communications amongst its members,
  - its composition: members drop, join and switch.

Our Interest...

- We'd like to:
  - Track and predict communication patterns amongst members,
  - rise of new COIs,
  - changes in existing COIs.
- We aim at profiling/characterising:
  - Some specific nodes and COIs,
  - and find nodes or COIs with similar behaviours.

Previous approaches

Amongst others...

- Cortes *et al.* [1] deal with a dynamic telephone network graph:
  - Focus on summarising data for easier daily storage and extraction.
  - Use of exponential smoothing.
  - Catch fraudsters based on calling patterns.
  - Have access to a list of fraudsters for comparison.
- Yu and Lambert [2] model time of day calling patterns.
  - Based on duration of calls, day of week and other variables.
  - Result: Each node or group of nodes are profiled under one of several calling patterns

Our Data...

- Origin (From).
- Destination (To).
- Time stamp.
- Some grouping labels.
- **N**O covariate data for us as in Yu and Lambert [2].
- **N**O known list for comparison for us as in Cortes *et al.* [1].
- Thousands of nodes.
- Varying number of transactions per day.

What are we doing?...

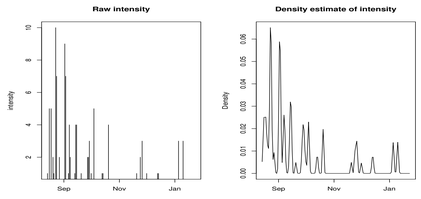
Relational measurements on static data

- We present some approaches to compare node and graph behaviour using connectedness of nodes and similarity measures based on number of calls (intensity).
- COIs can be formed based on the frequency of contact between nodes (Connectedness) or based on the pattern of calls made even if these calls don't involve contact between two members of a COI.
- These first approaches are static, that is, taken over the complete time period as a whole.
- **Connectedness:** Based on the communications shared by two nodes we assign connectedness scores to every pair.

$$Con(A,B) = \frac{X_{AB}X_{BA}}{\sum_{k=1}^{n}(X_{Ak}+X_{kA}+X_{Bk}+X_{kB})} \tag{1}$$

where $X_{AB} = \#$ of calls from node A to node B and $0 \leq Con(A,B) = Con(B,A) \leq 0.5$

- **Similarity:** Based on call density curves estimated by kernel density smoothers on the number of daily communications.
- Densities spread out the effect of the discretely observed intensity of communications.
  - Model communication intensity (ignoring who originated the communication):
    t = time (in days),
    $Y_i(t)$ = number of communications involving node i,
    $\mu_i(t) = E(Y_i(t))$ = Intensity function.
  - Estimate $\mu_i(t)$ with $Y_i(t)$ via density estimation with triangular kernels.



  - Density for node i is

$$\mu_i(t) = \frac{1}{nh}\sum_{i}^{n} K(\frac{x-t}{h})$$

where K is the Kernel function with bandwith h. Then

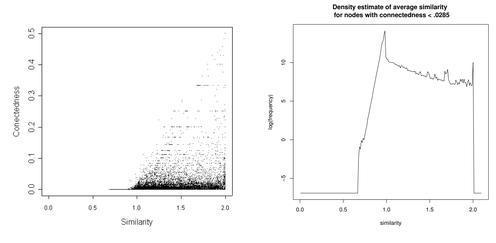
$$0 \leq Sim(A,B) = 1 + cor(\mu_A, \mu_B) \leq 2 \tag{2}$$

where $cor(A,B)$ is Pearson's correlation coefficient between the two density vectors. This similarity measurement is invariant to scaling.

- Search for grouping/clustering models based on connectedness and similarity and compare to known grouping labels.

Some findings

- Comparisons based on connectedness and similarity
  - Communications are sparse. Hence high similarity values (> 1.5) even if low connectedness
    - over 75% nodes make 6 calls or less in a 180 day period



  - If nodes i and j communicate once but not to each other they have a similar pattern.
- Few nodes ($\sim 10\%$) do both, originate and receive communications.
- Those who only generate communications generate 9 on average.
- Those who only receive communications receive 29 on average.

A bit on probability models...

- Model directly on the intensity:
  - Distribution for $Y_i(t)$: Poisson ($\mu_i(t)$) as we have counts of communications per time period.
  - Model $\mu_i(t)$ as a piece wise constant function.
- Model on node or group's communication pattern.
  - Let $n_{Ai_k} = \#$ of communications from node A to node i_k of group k,
  - $n_{Ai_k} \sim$ multinomial $(p_{A1_k}, p_{A2_k}, \cdots p_{AI_k})$, where $p_{AA} = 0$ as we have no self communications.
  - The p_{Ai_k}'s are realisations from a *Dirichlet* distribution,
  - this constitutes node A's profile with respect to group k, say P_k.
  - Overall profile of A is a mixture model: $\pi_{1A}P_1 + \pi_{2A}P_2 + \cdots \pi_{KA}P_K$.
  - To identify or monitor groups we need to estimate π_{iA} in a fixed time period and track changes.

How would we estimate?

- EM algorithm, MCMC come to mind.

What about the dynamics?

Proposed approaches and ideas:

- In order to detect changes in groups through time:
  - Bin data in fixed time period (width) windows.
  - Fit model(s) on these.
  - Slide windows and track changes in model(s).
- More complex models
  - Time varying p's in *Dirichlet* distribution?
  - Different *Dirichlet* distributions at different time periods?
- Testing on $\mu(t)$ to detect changes
  - Sequential: Modifications of Wald's test.
  - Assume time windows as independent observation times and test estimate from current window against estimate from previous window, rather than sliding the window.

Challenges:

- Data volume
  - Tens of thousands of nodes.
  - Similarity and connectedness indexes are calculated pairwise: Must calculate in the order of n^2 values for each.
  - Sparse connectivity.
- Data quality:
  - Different data streams.
  - Lack information on activity of data streams.
- Running time, need to set up process to run in parallel.

- Contacts:
  - alberto.nettel-aguirre@acadiau.ca
  - hugh.chipman@acadiau.ca

References

[1] Corinna Cortes, Daryl Pregibon, and Chris Volinsky. Computational methods for dynamic graphs. *Journal of Computational and Graphical Statistics*, 12(4):950–970, 2003.

[2] Yan Yu and Diane Lambert. Fitting trees to functional data, with an application to time-of-day patterns. *Journal of Computational and Graphical Statistics*, 8:749–762, 1999.