Tracking Communities of Spammers by Evolutionary Clustering

Kevin Xu[1], Mark Kliger[2], Alfred O. Hero III[1]

[1]University of Michigan, Ann Arbor, MI, USA

[2]Medasense Biometrics, Ofakim, Israel

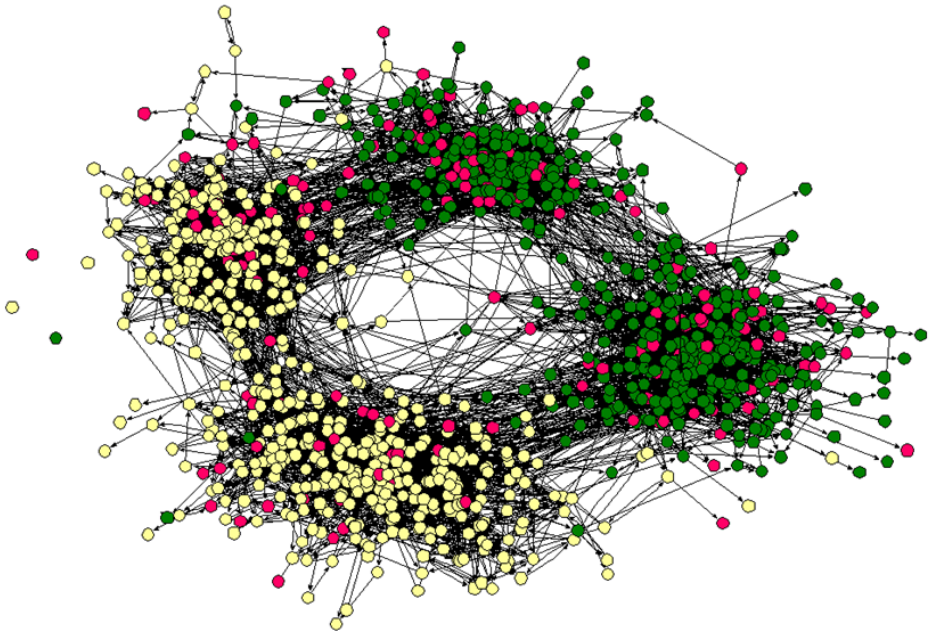
Presented by Mark Kliger

Outline

Outline

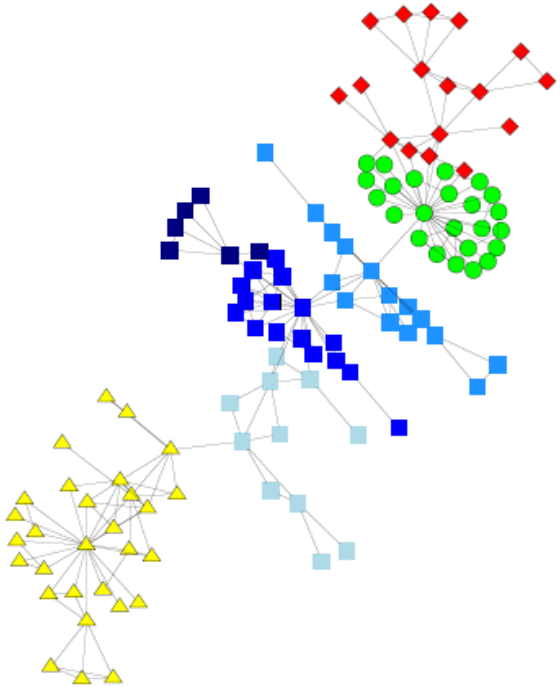
Communities in Social Networks

School friendships



Moody, 2001

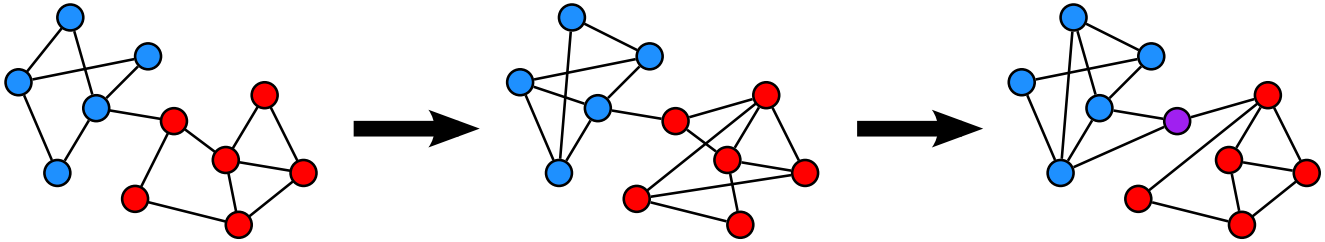
Scientific collaborations



Girvan and Newman, 2002

- Detecting Communities in Social Networks is a popular subject.
- Various algorithms
  - Leskovec et al. (2010) for empirical comparison of different algorithm

Dynamic Social Networks



- Almost ALL social networks are changing in time.
- Objectives of the study:To track changes in community structure over time
- Trigger project: To reveal communities of spammers!

Stages of SPAMming process

Legal

Illegal (almost...)

- First Stage: Harvesting - mass acquisition of email addresses using harvesters (bots, crawlers, web-spiders, etc.)
- Second Stage: Spamming - sending large amounts of spam emails using spam servers
- Observation: Spammers conceal their identity to a lesser degree when harvesting (Prince:CEAS2005)
- Spammers might be associated with their harvesting means

www.projecthoneypot.org

- Distributed network of decoy web pages - “honey pots”.

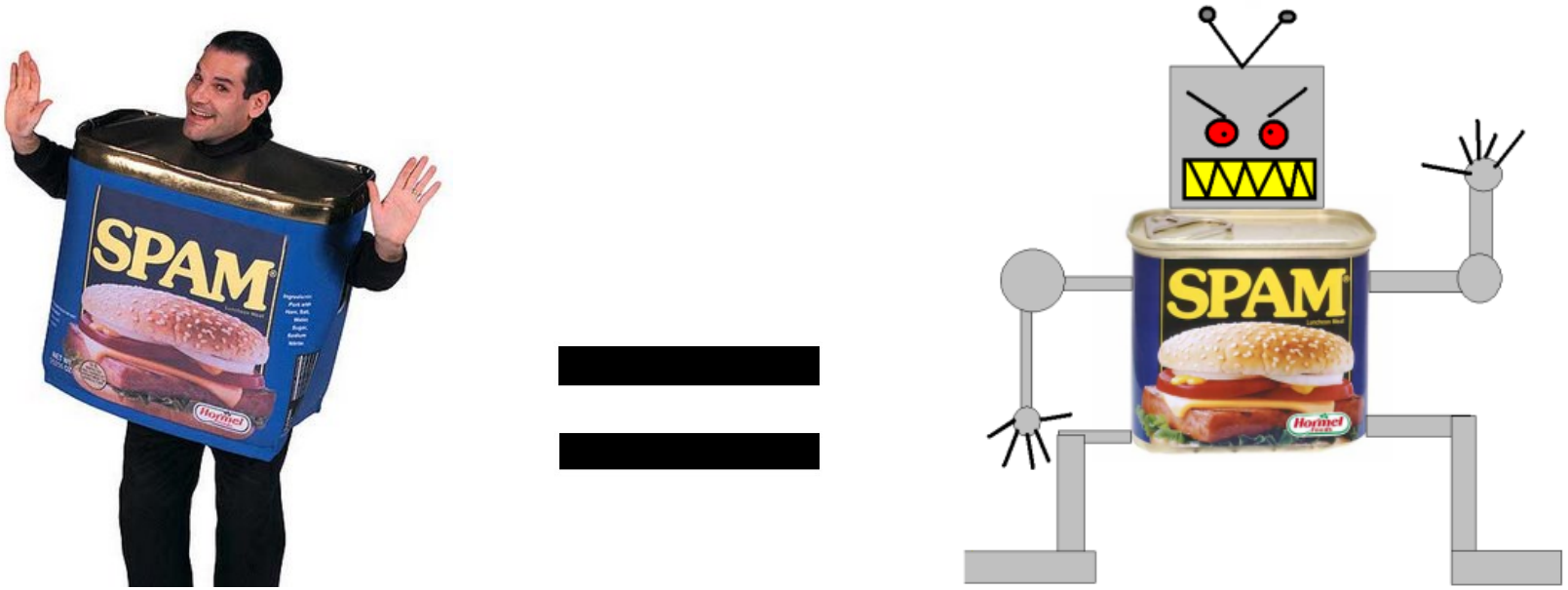
SPECIAL LICENSE RESTRICTIONS FOR NON-HUMAN VISITO RS

Special restrict ions on a visitor's license to access the Websit e apply to Non-Human Visitors. Non-Human Visitors i nclude, but are not limi ted to, web spiders, bots, indexers, robots, crawlers, harvesters, or any other comp uter programs designed to access, read, compile or gather content from the We bsite automati cally.

Email ad dresses on this site are considered prop rietary intellectual pro perty. It is recognized that these email addres ses are provided for human visit ors alone. You acknowledge and agree that each email address the Website con tains has a va lue not less than US $50. You further agree that the compilation , storage, and/or distribu tion of these ad dresses substantially diminishes the value of these addresses. Intentional collection, harvesting, gathering, and/or storing this Website's email addresses is recogn ized as a violation of t his agreement and ex pressly prohibit ed.

- Honey Pot: text of a legal document with trap email address embedded inside HTML code

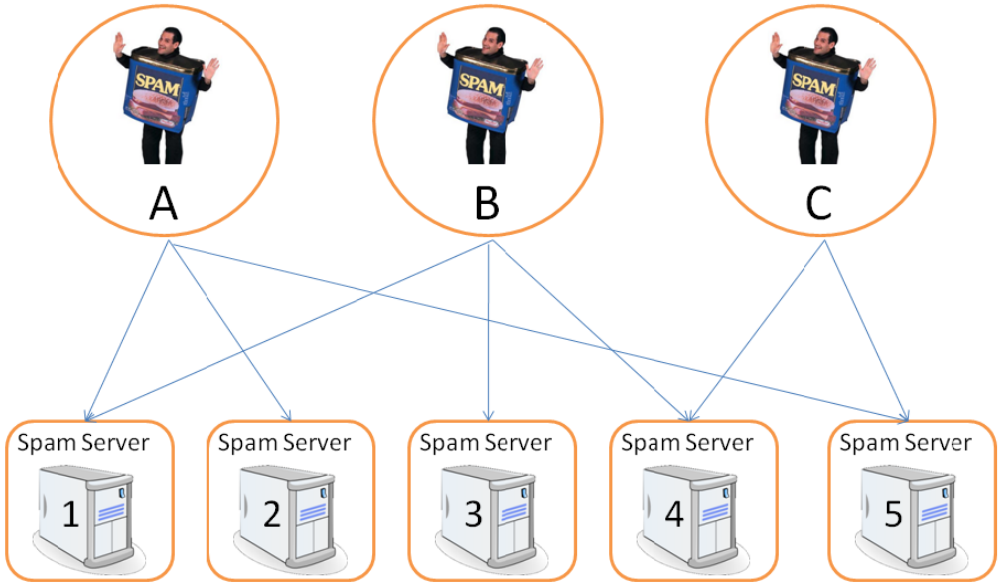
Tracking Spammers



- Non-human visitor (bot, crawler, spider, harvester i.e. spammers) hit the honey pot and collect trap email address.
- Spammer IP address is stamped and tracked
- Unique email address generated each visit. Email addresses and all received messages associated with a single spammer. All messages are spam.

Network of Spammers

- How do we characterize social networks and communities?
  - Social interactions between members
  - Sharing resources between members
  - Similarity in members' behaviors



- Ties between spammers by shared spam servers

Strength of Ties

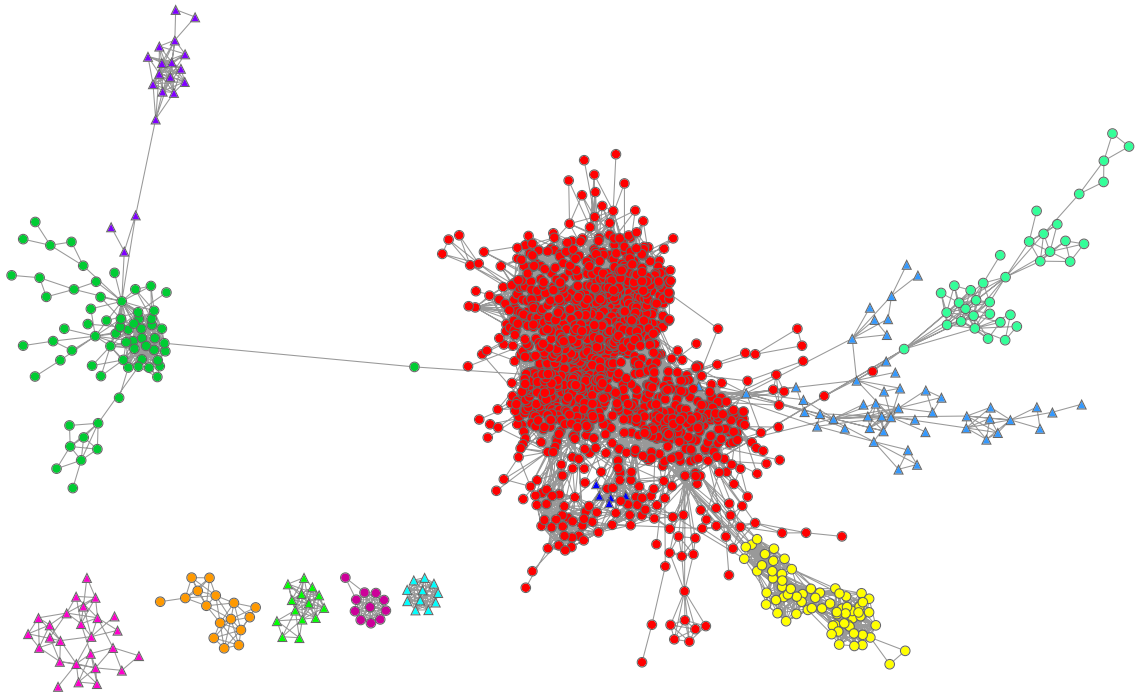
- Connect spammers by similarity in spam server usage
- Coincidence matrix H^t between spammers and spam servers at time point t:

$$H^t = \left[\frac{p_{ij}^t}{e_i^t} \right]_{i,j=1}^{M,N}$$

- p_{ij}^t: number of emails sent by spammer i using spam server j during time interval t
- e_i^t: total number of email addresses collected by spammer i up to time t
- Network of spammers is represented by dot product affinity matrix:

$$W^t = H^t (H^t)^T$$

Static Communities of Spammers (Xu et al, 2009)



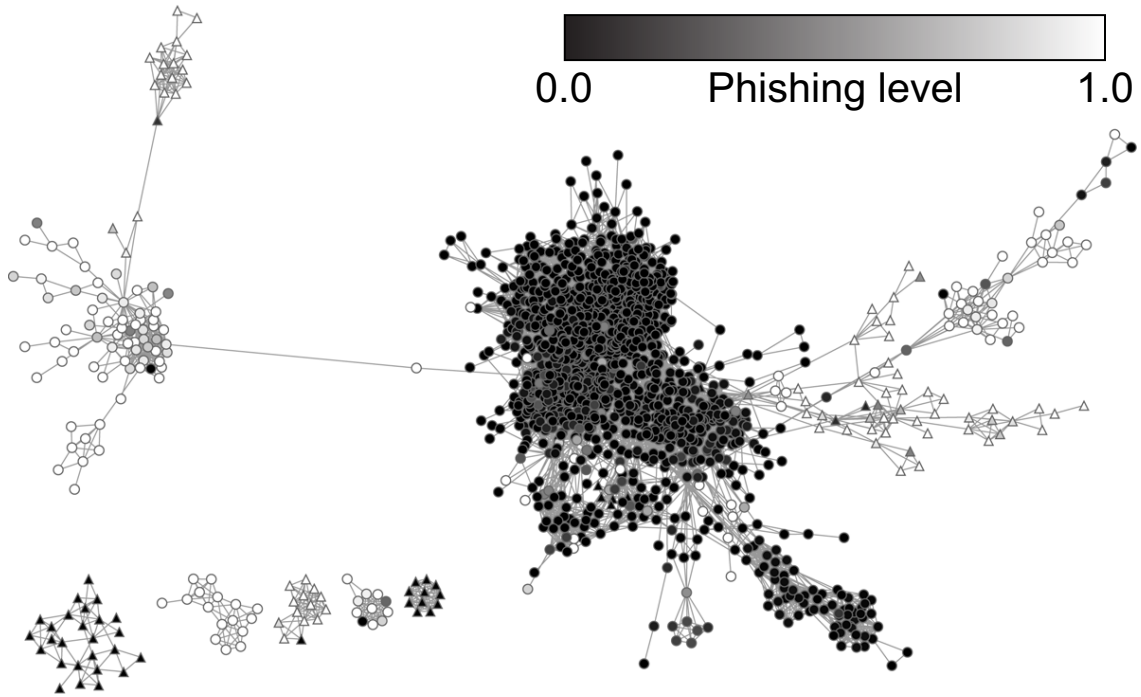
Oct. 2006

- Multiclass Spectral Clustering (Yu and Shi, 2003): Relaxation of

$$\max_X \frac{1}{K} \sum_{i=1}^{K} \frac{\mathbf{x_i}^T W^t \mathbf{x_i}}{\mathbf{x_i}^T D^t \mathbf{x_i}}$$

s.t. $X = [\mathbf{x_1} \cdots \mathbf{x_K}] \in \{0,1\}^{M \times K}; X 1_K = 1_M; D^t = diag(W^t 1_M)$

Static Communities of Spammers (Xu et al, 2009)



Oct. 2006

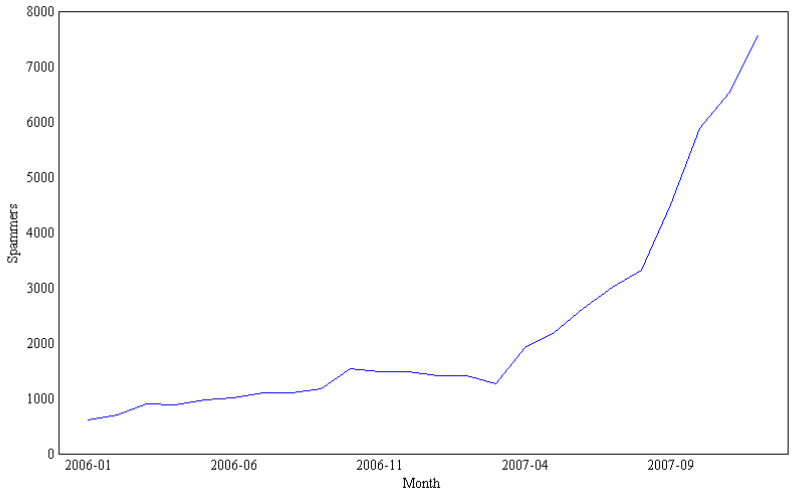
- Validation by phishing level spammer

$$\text{Phishing level} = \frac{\text{\# of phishing emails sent}}{\text{total \# of emails sent}}$$

- Email classified as phishing email if subject contains common phishing word (e-Bay, PayPal, Chase, passport, login, etc.)

Dynamic Network of Spammers

- Project Honey Pot has grown exponentially with time



- As of June 2010
  - 45 million trap email addresses monitored
  - 67 million spam servers identified
  - more then billion spam messages received
  - 79 thousands spammers identified

- Our goal: to identify and track communities of spammers over time

Outline

Community detection in dynamic social networks

- Ignore history and cluster only current data
  - Clustering results are unstable
- Evolutionary Clustering
  - Incorporate both past and present data

$$\bar{W}^t = \alpha^t \bar{W}^{t-1} + (1 - \alpha^t) W^t$$
$$(\bar{W}^0 = W^0)$$

- Forgetting factor α^t controls the amount of smoothing
- Evolutionary Spectral Clustering - spectral clustering with $\bar{W}^t$ (Chie et al, 2007)
- How to select α^t?

Optimal forgetting factor

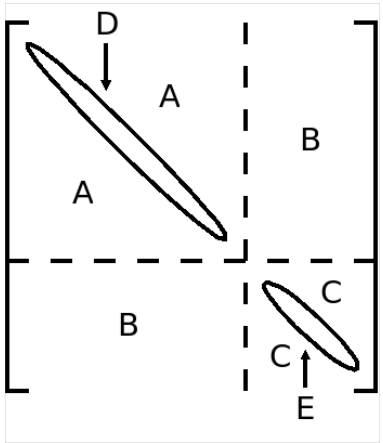
- Borrowing ideas from Shrinkage Estimation of Covariance matrices (Ledoit and Wolf, 2003)
- Assume that: True affinity matrix at any given time t to be the expected affinity matrix $E(W^t)$.
- Optimum α^t in Minimum Mean Square Error sense (MSE)

$$(\alpha^t)^* = \underset{\alpha \in [0,1]}{\operatorname{argmin}} \, \mathrm{E}\left[\| \alpha \bar{W}^{t-1} + (1-\alpha) W^t - E(W^t) \|_F^2 \right]$$

$$= \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{var}(w_{ij}^t)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ [\bar{w}_{ij}^{t-1} - \mathrm{E}(w_{ij}^t)]^2 + \operatorname{var}(w_{ij}^t) \right\}}$$

Oracle is on vacation....

- $(\alpha^t)^*$ is not implementable because it requires knowledge of the mean and variance of the entries of W^t
- Replace unknowns with sample statistics
- Sample mean and sample variance of W^t are dependent on clustering structure of G^t



- We don't know which samples belong to which cluster.

This is the goal of clustering!

Iterative estimation component memberships and α^t

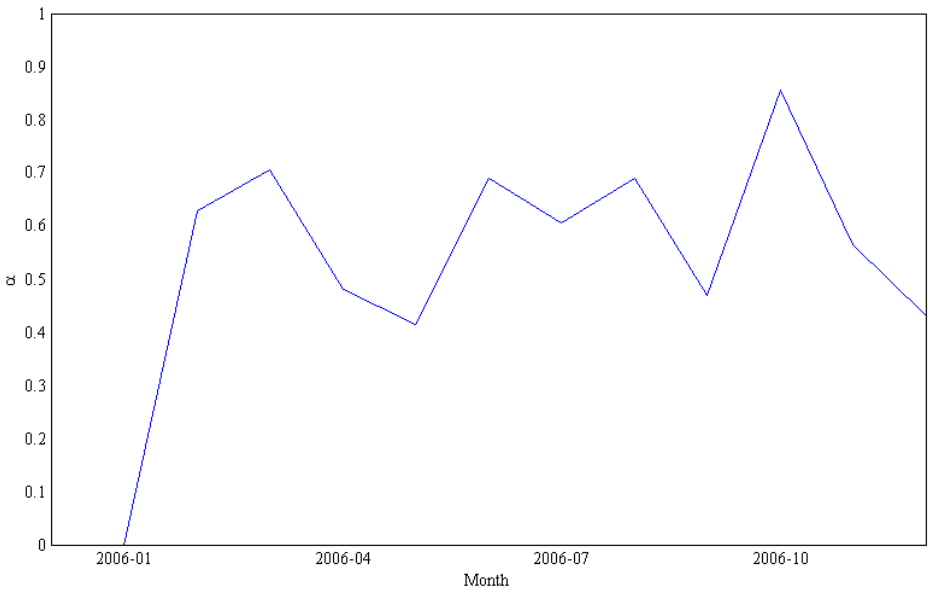
1. Fix component memberships to be the most recent cluster memberships
2. Estimate sample mean and variance of W^t by summing over each cluster.
3. Calculate α^t and $\bar{W}^t$
4. Fix $\bar{W}^t$, and run clustering algorithm to obtain new cluster memberships
5. Repeat entire procedure (until α^t converges...)

- We haven't proved that α^t converges but empirically it "converges" after only a handful of iterations.

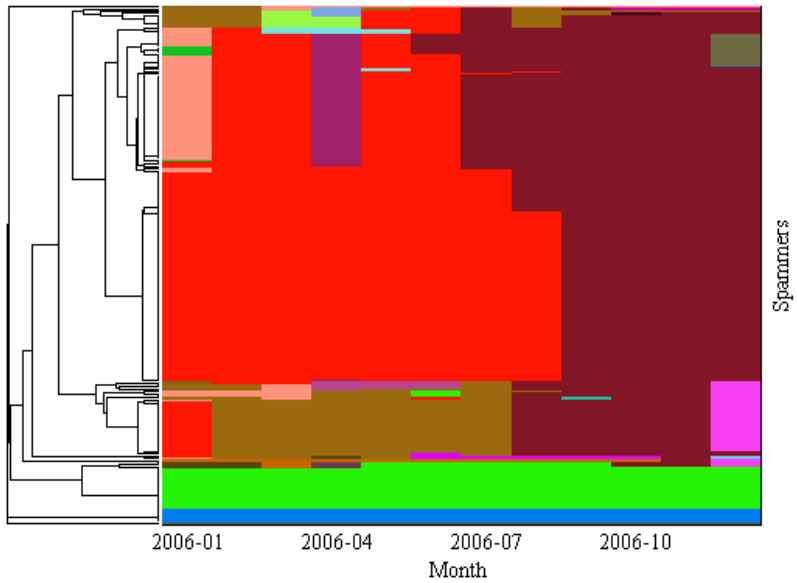
Outline

Estimation of α^t (2006 monthly)

Estimated forgetting factor α^t

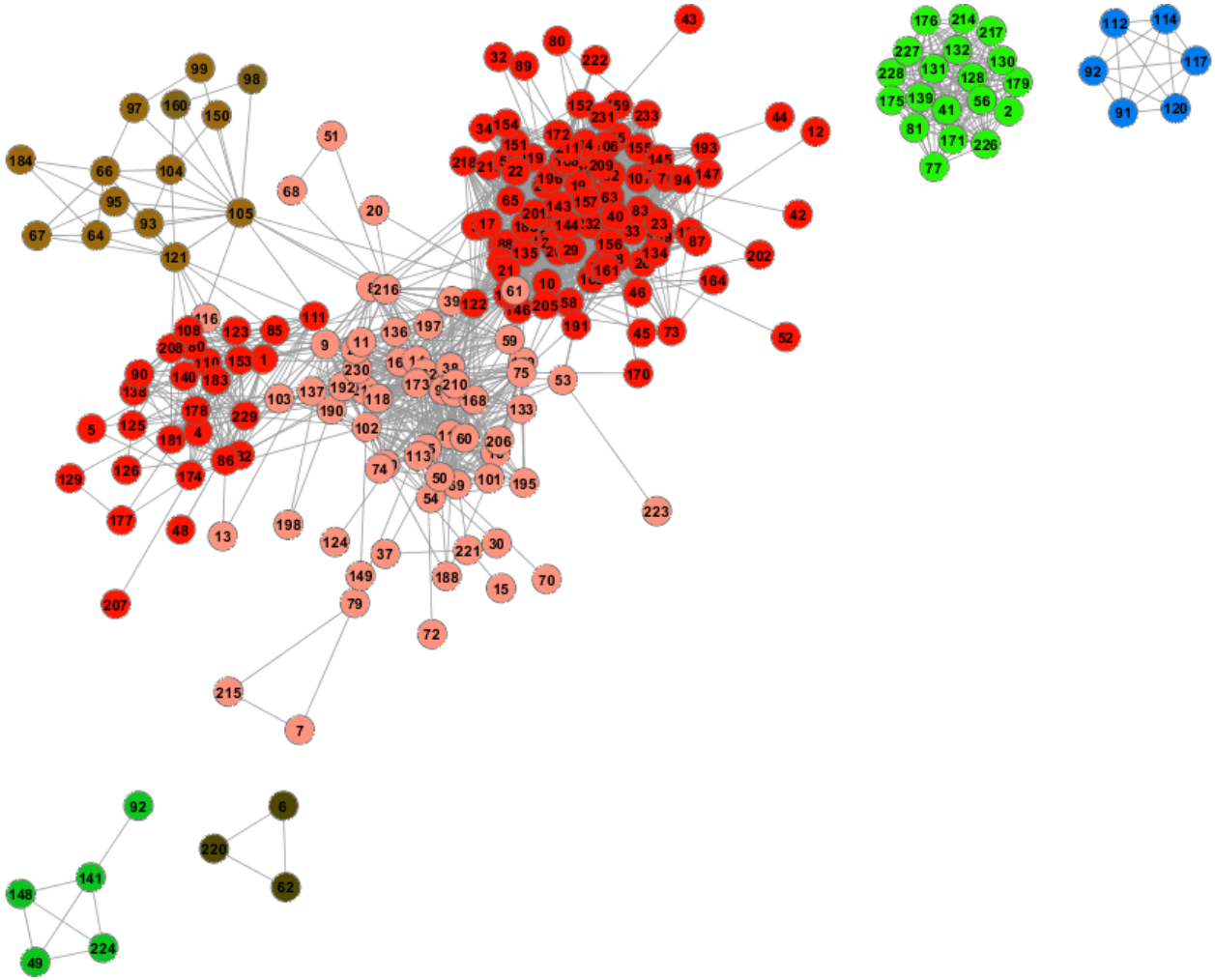


Community memberships (240)



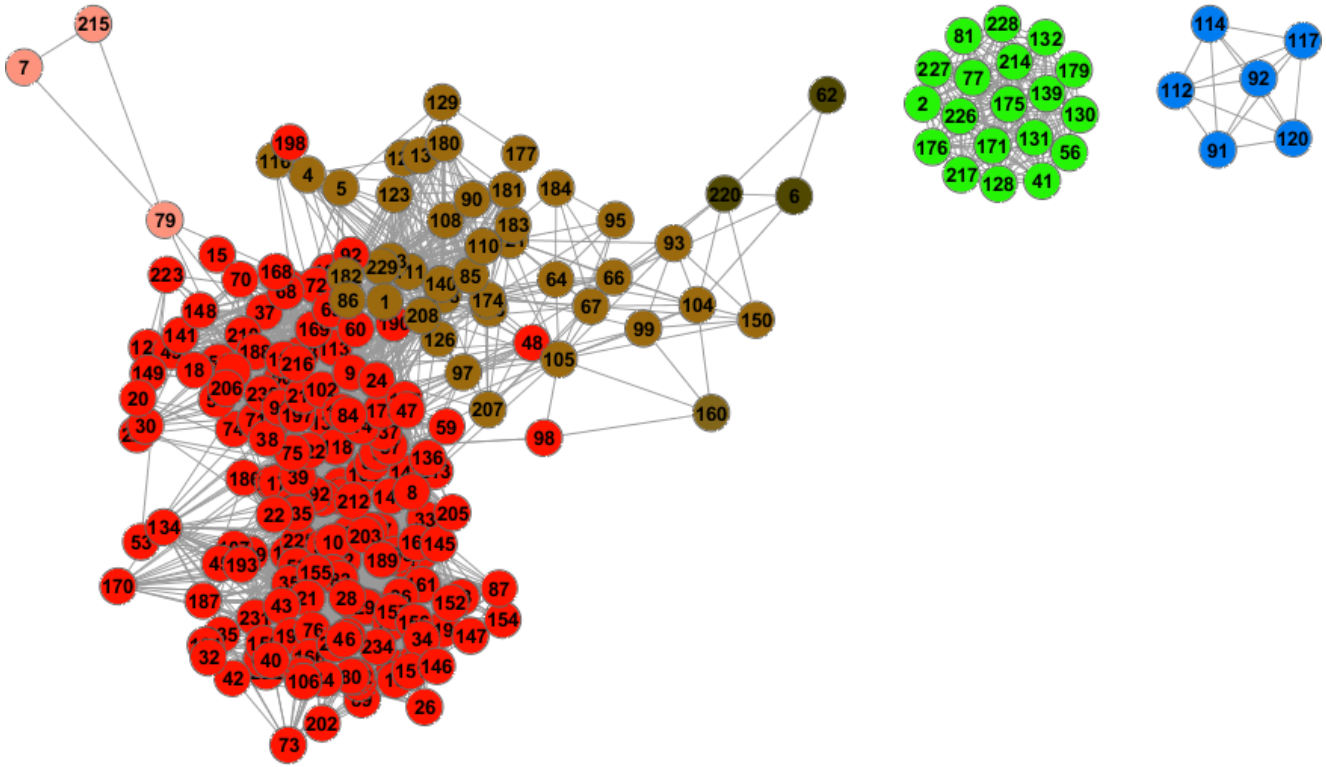
- α^t changes around January, April, September, and December, suggesting changes in the community structure during these months
- No validation is available
- Difficult to visualize dynamic network

Preliminary Results (2006 monthly) - 240 spammers



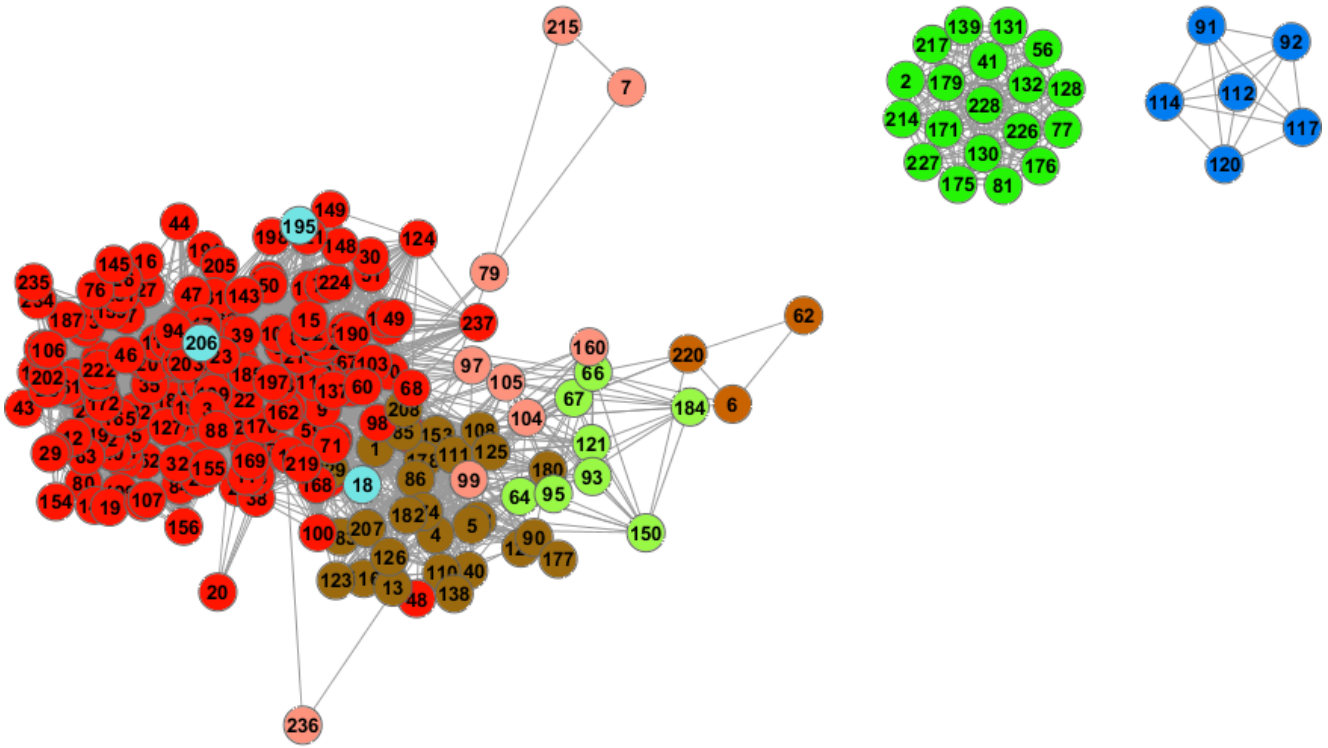
01.2006

Preliminary Results (2006 monthly) - 240 spammers



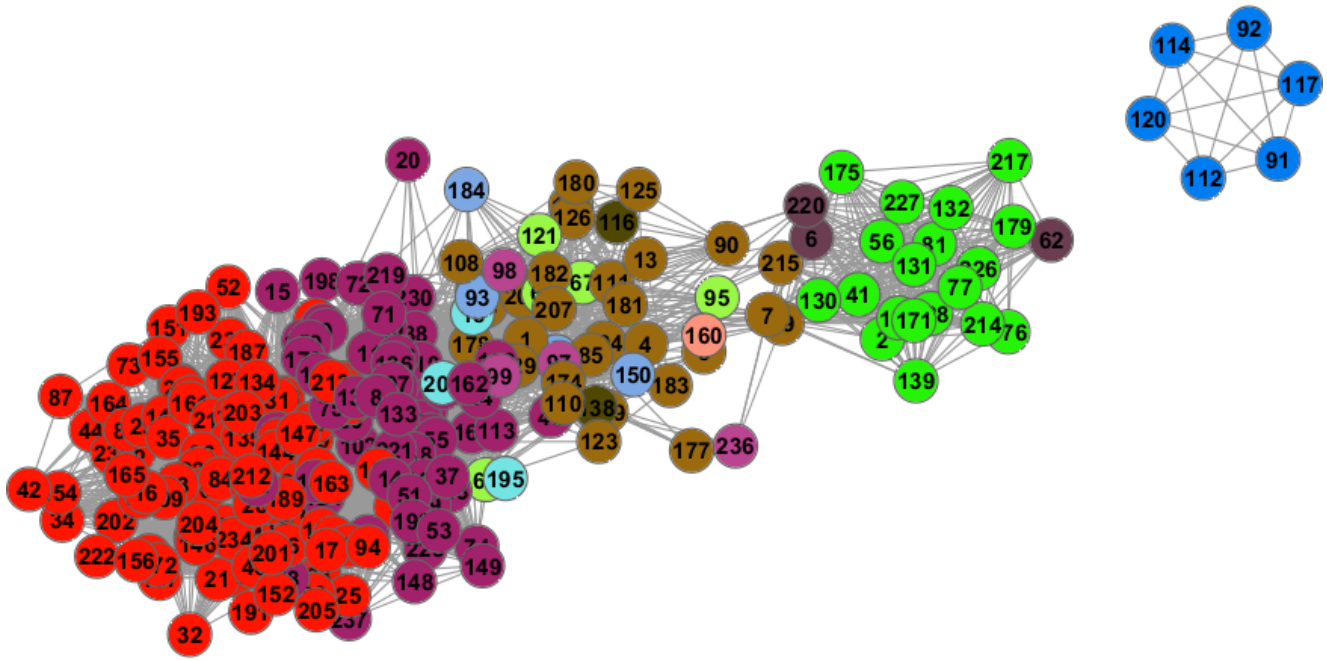
02.2006

Preliminary Results (2006 monthly) - 240 spammers



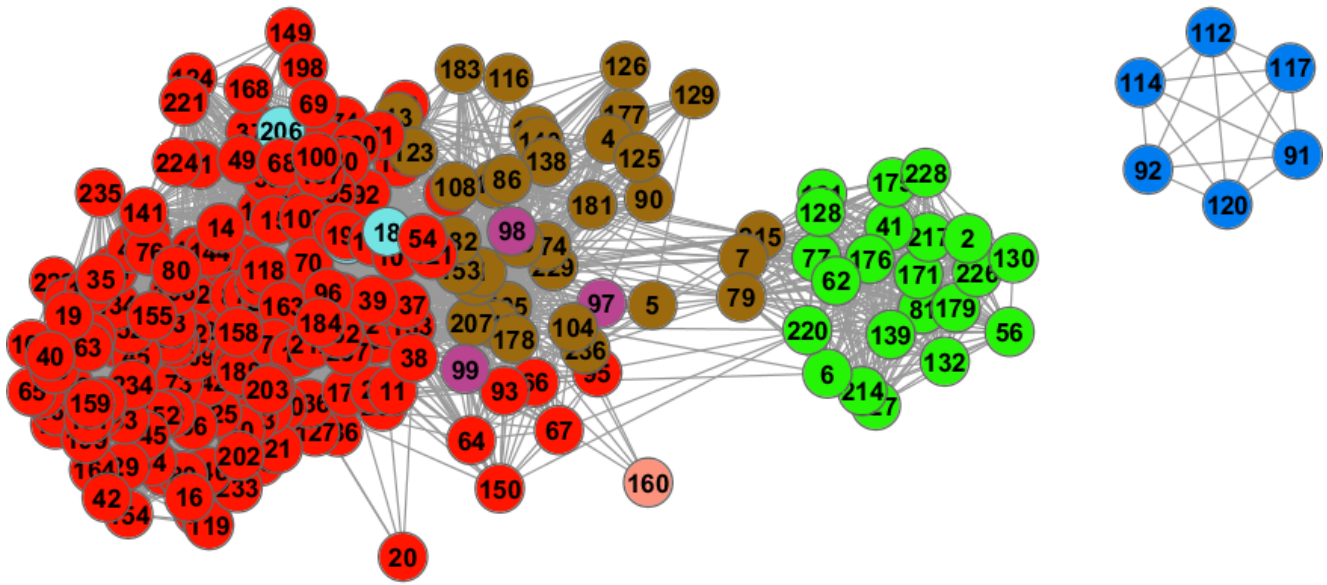
03.2006

Preliminary Results (2006 monthly) - 240 spammers



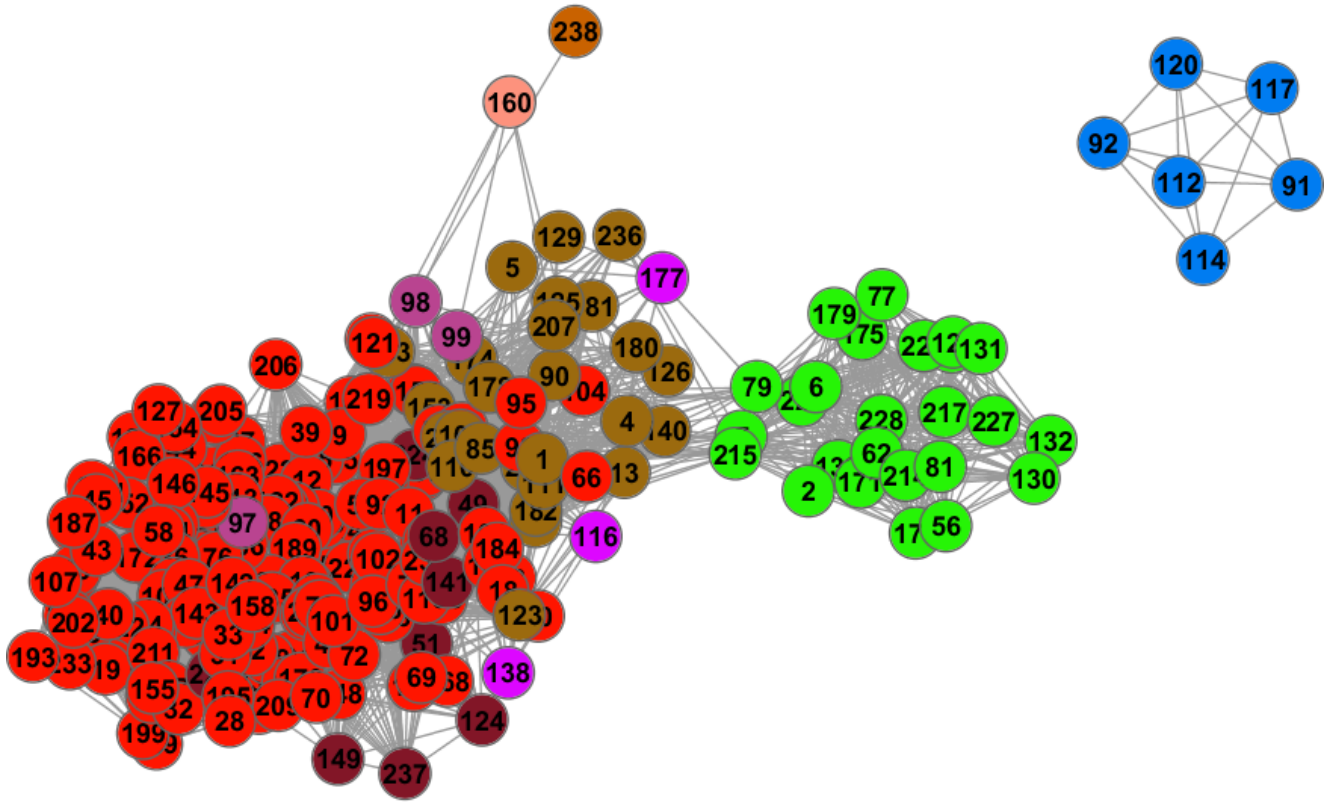
04.2006

Preliminary Results (2006 monthly) - 240 spammers



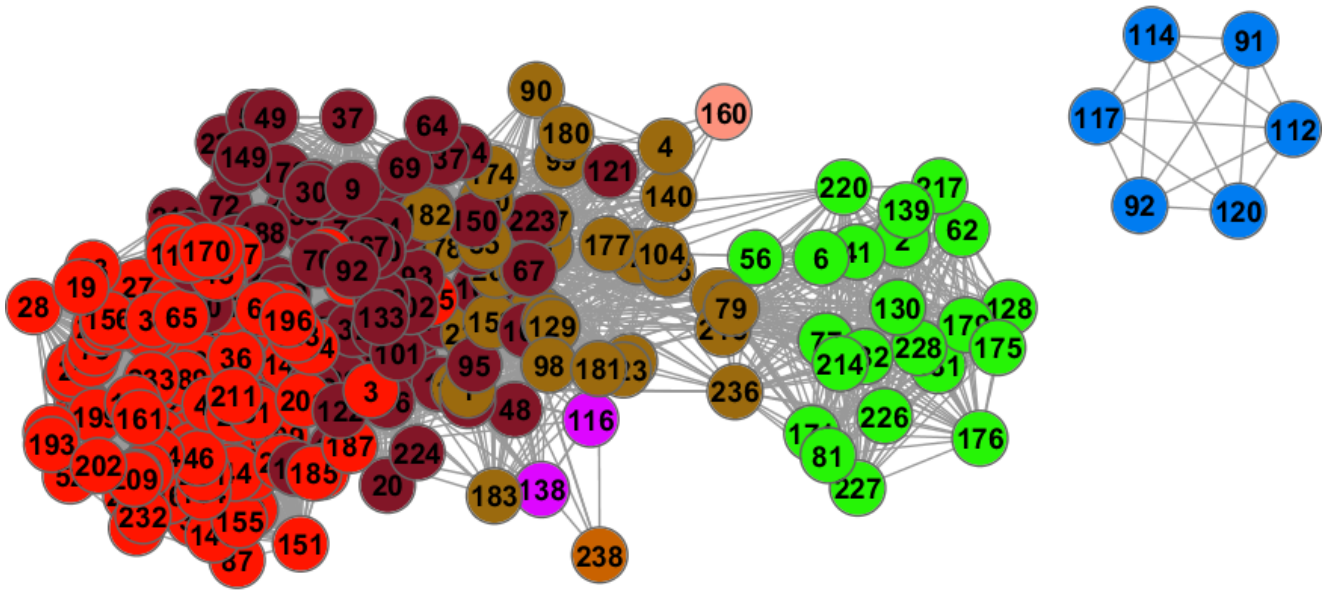
05.2006

Preliminary Results (2006 monthly) - 240 spammers



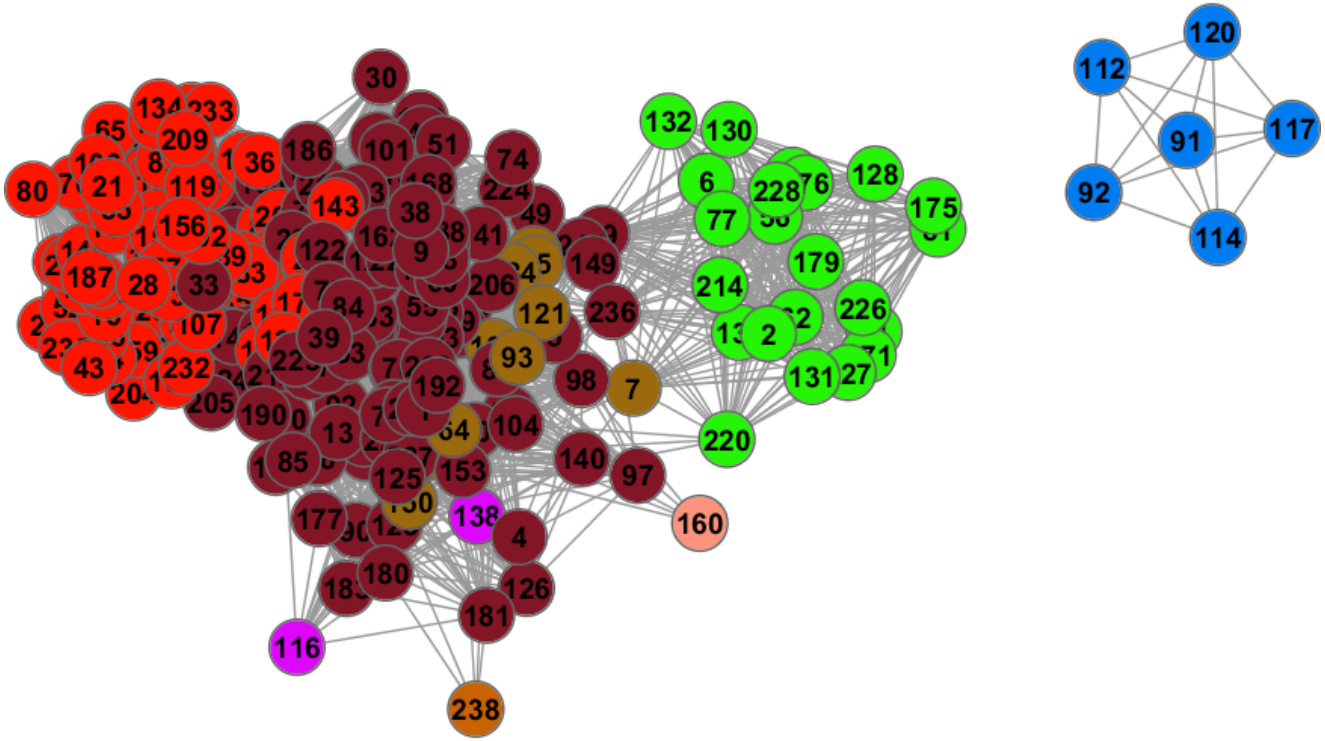
06.2006

Preliminary Results (2006 monthly) - 240 spammers



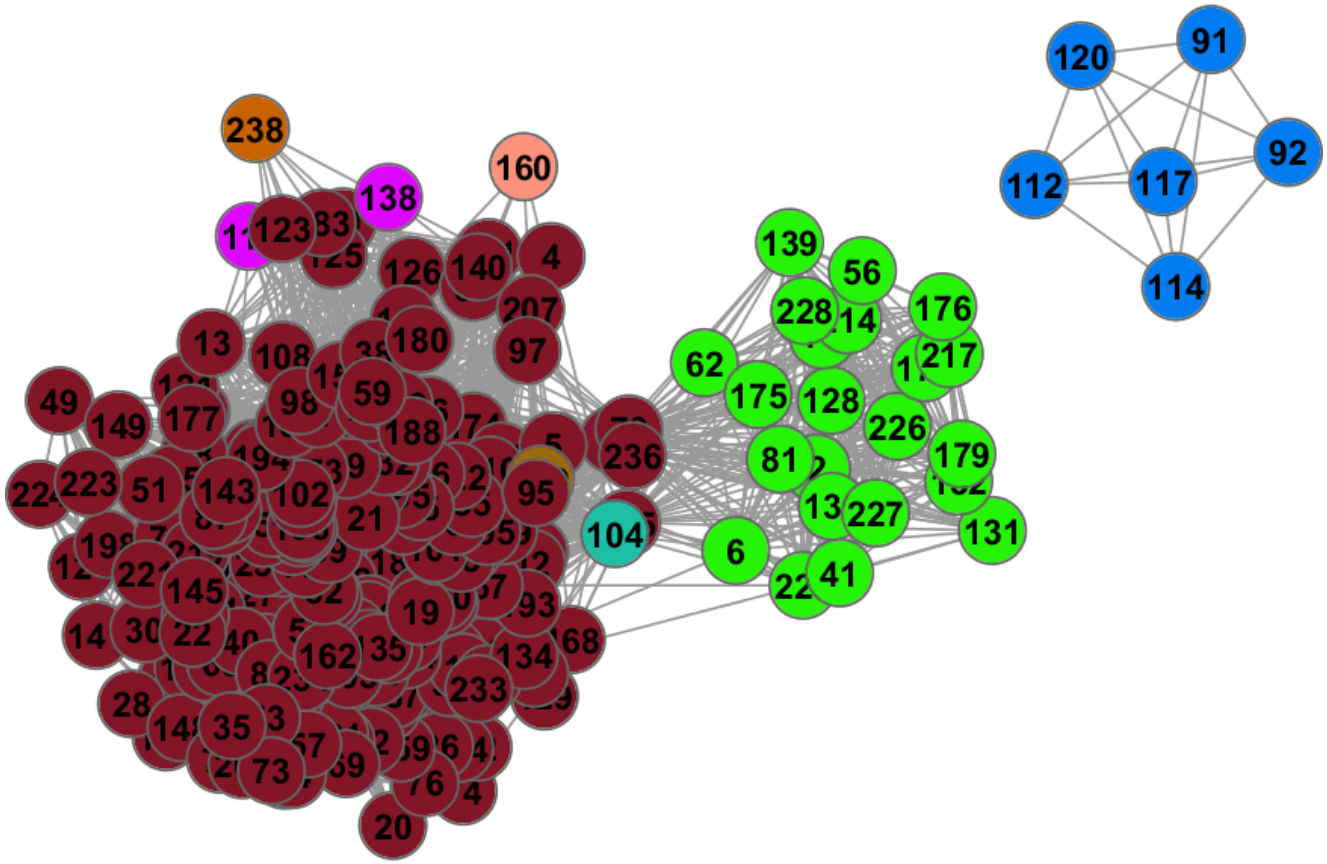
07.2006

Preliminary Results (2006 monthly) - 240 spammers



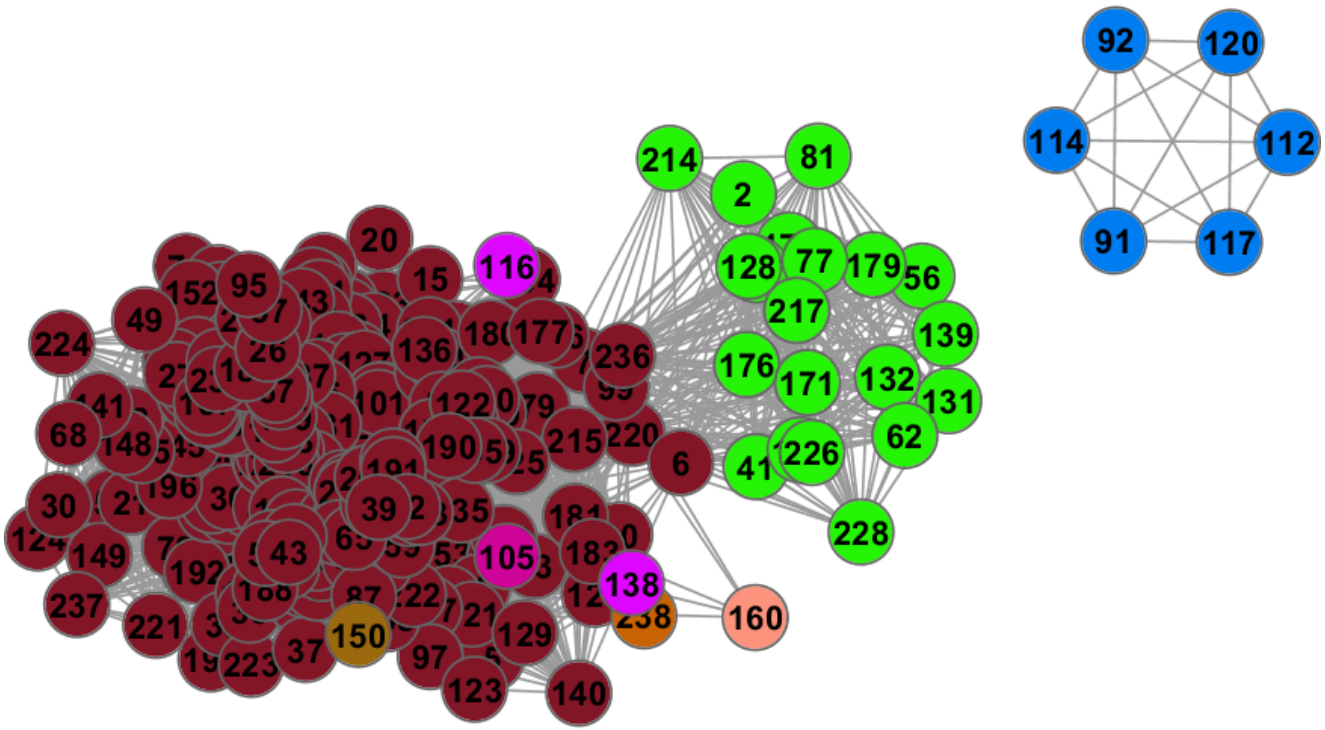
08.2006

Preliminary Results (2006 monthly) - 240 spammers



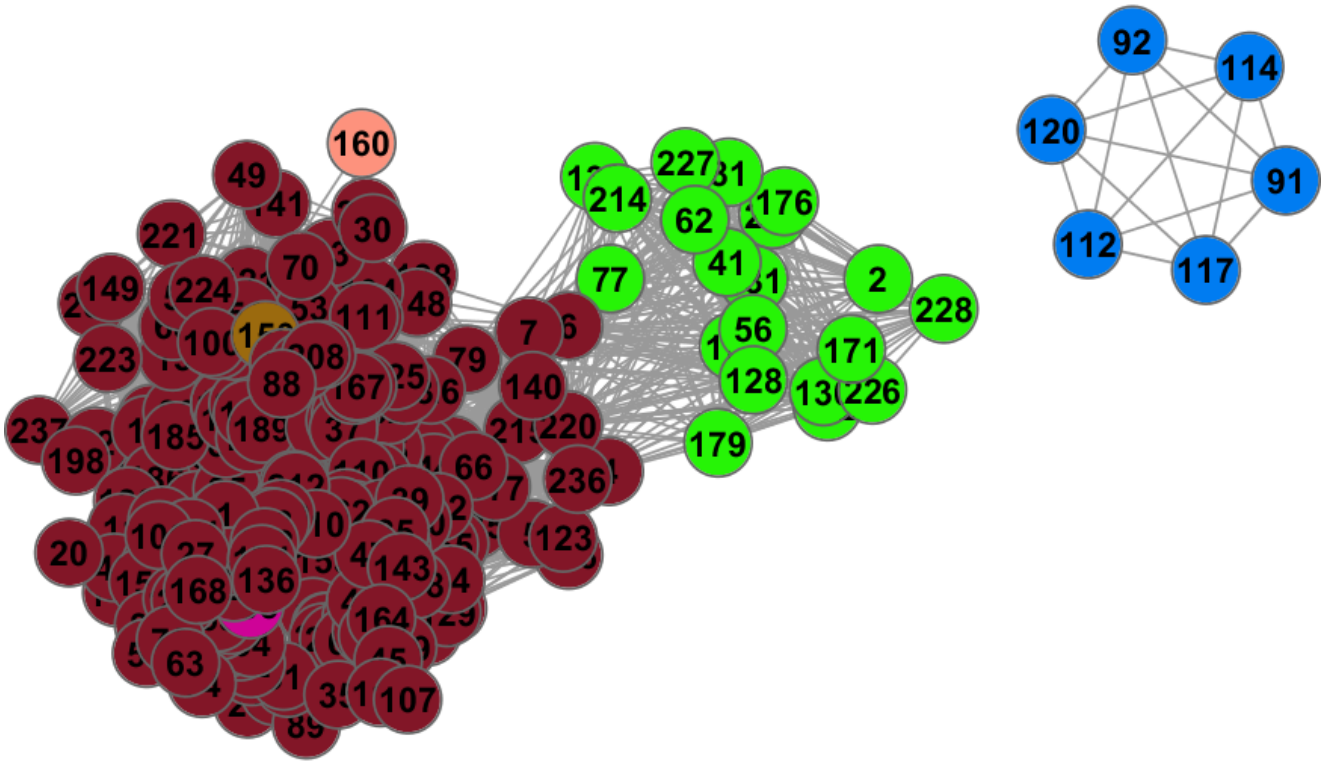
09.2006

Preliminary Results (2006 monthly) - 240 spammers



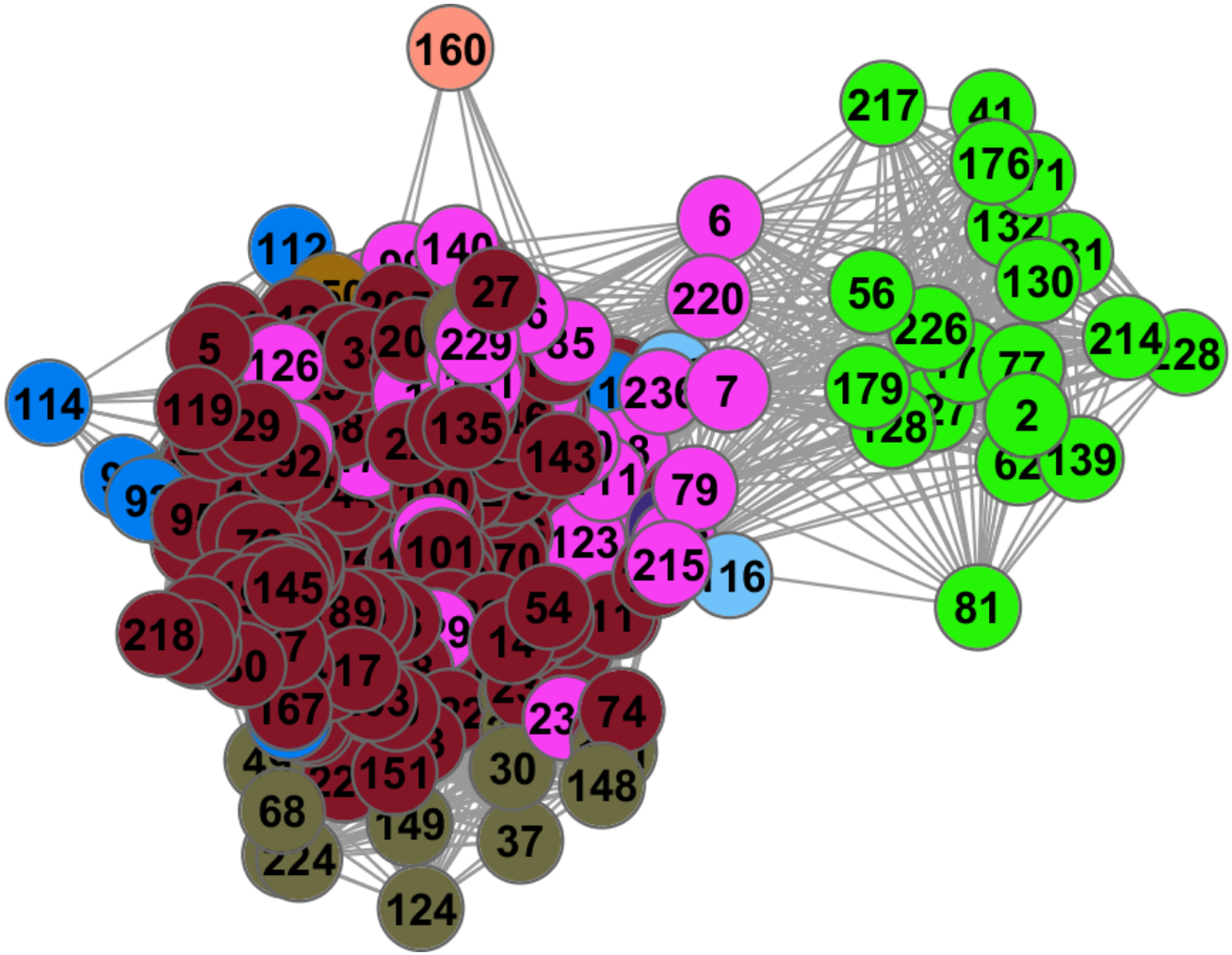
10.2006

Preliminary Results (2006 monthly) - 240 spammers



11.2006

Preliminary Results (2006 monthly) - 240 spammers



12.2006

Outline

Challenges

- How to validate a clustering result in unlabeled social network?
  - Indirect validation: compare α^t with times of known major events or change points, if such information is available
- Properly choosing number of communities
  - EigenGap heuristic (von Luxburg, 2007) on $\bar{W}^t$
  - One would expect that the number of communities, much like the community memberships, should vary smoothly with time
- Visualization of dynamic network?
  - Force-directed layout (we use Cytoscape) is sucks for visualization of dynamic networks.
- Your opinion how to analyze and validate this data will be much appreciated!

- Thanks to Unspam Technologies for providing data from Project Honeypot
- This work was partially supported by:
  - National Science Foundation grant CCF 0830490
  - Office of Naval Research grant N00014-08-1-1065.

Questions?

References

1. M. Girvan and M. E. J. Newman, “Community Structure in Social and Biological Networks,” *Proc. National Academy of Sciences* (2002).
2. J. Moody, “Race, School Integration, and Friendship Segmentation in America,” *American Journal of Sociology* (2001).
3. J. Leskovec, K. Lang, M. Mahoney, “Empirical Comparison of Algorithms for Network Community Detection," ACM WWW International conference on World Wide Web (WWW), 2010.
4. M. Prince et al., “Understanding How Spammers Steal Your E-Mail Address: An Analysis of the First Six Months of Data from Project Honey Pot,” *2nd Conference on Email and Anti-Spam* (2005).
5. U. von Luxburg, “A Tutorial on Spectral Clustering,” *Statistics and Computing*, (2007).
6. K. S. Xu, M. Kliger, Y. Chen, P.J. Woolf and A.O. Hero, “Revealing Social Networks of Spammers Through Spectral Clustering,” IEEE ICC 2009.
7. S. Yu and J. Shi, “Multiclass Spectral Clustering,” *9th IEEE International Conference on Computer Vision* (2003).
8. Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng, ”Evolutionary spectral clustering by incorporating temporal smoothness,” KDD 2007.
9. O. Ledoit and M. Wolf, ”Improved estimation of the covariance matrix of stock returns with an application to portfolio selection,” Journal of Empirical Finance, 2003.

The End!