A Network Model for Dynamic Textual Communications with Application to Government Email Corpora

Bomin Kim[1] Aaron Schein[3]
Bruce Desmarais [1] Hanna Wallach[2,3]

[1] The Pennsylvania State University

[2] Microsoft Research NYC

[3] University of Massachusetts Amherst

June 8, 2017

Interaction-Partitioned Topic Model (IPTM)

- Probablistic model for time-stamped textual communications (e.g. emails, cosponsorship of bills, international sanctions)
- Integration of two generative models:
  - Latent Dirichlet allocation (LDA) for topic-based contents
  - Dynamic exponential random graph model (ERGM) for ties
- IPTM assigns each topic to an "interaction pattern," which is governed by a set of dynamic network features

"who communicates with whom about what, and when?"

Content Generating Process: LDA (Blei et al., 2003)

- For each topic $k = 1, ..., K$:

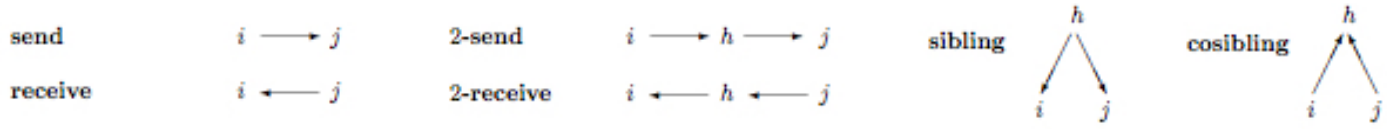
  1. Topic-word distribution $\boldsymbol{\phi}^{(k)} \sim$ Dirichlet$(\beta, \mathbf{u})$
     - A topic k is characterized by a discrete distribution over V word types with probability vector $\boldsymbol{\phi}^{(k)}$.
  2. Topic-IP distribution $c_k \sim$ Uniform$(1, C)$
     - Each topic is associated with a single interaction pattern.

- For each document $d = 1, ..., D$:

  3-1. Document-topic distribution $\boldsymbol{\theta}^{(d)} \sim$ Dirichlet$(\alpha, \boldsymbol{m})$
     - A document d is characterized by a discrete distribution over K topics with probability vector $\boldsymbol{\theta}^{(d)}$.

  3-2. For each word in a document $n = 1$ to $N^{(d)}$:
     (a) Choose a topic $z_n^{(d)} \sim$ Multinomial$(\boldsymbol{\theta}^{(d)})$
     (b) Choose a word $w_n^{(d)} \sim$ Multinomial$(\boldsymbol{\phi}^{(z_n^{(d)})})$

Dynamic Network Features (Perry and Wolfe, 2012)

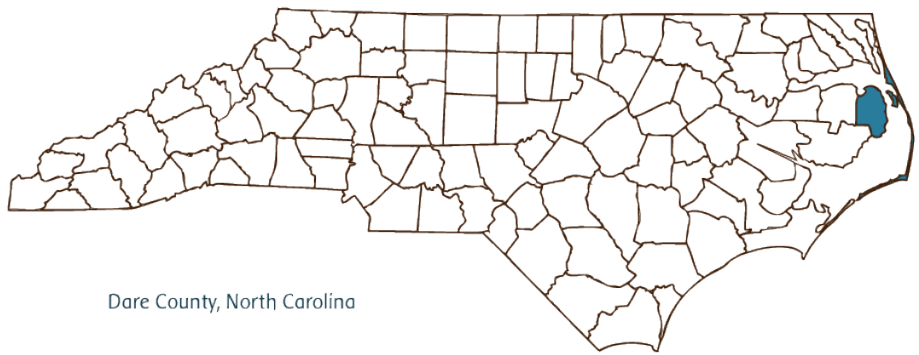


Tie Generating Process

Inference - Pseudocode

Data: North Carolina Dare county email data

- $D = 1456$ emails between $A = 27$ county government managers, covering 2 month periods (October 1 - November 30) in 2013



Effect of Hurricane Sandy

IPTM Result

Conclusion

- Joint modeling of ties (sender, receiver, time) and contents
- Allowance of multicast – multiple senders and/or receivers
- Possible application to