Learning a social network from text data

Xiaoyi Yang
Advisor: Nynke Niezink & Rebecca Nugent
Department of Statistics & Data Science

In collaboration with Christopher Warren
Department of English

Carnegie Mellon University

May 29, 2020

Learning modern social network

Modern social networks can be learned from various sources

Survey:
List your 20 best friends

Co-authorship:
Who are working together?

Warren, C. N., Shore, D., Otis, J., Wang, L., Finegold, M., & Shalizi, C. (2016). Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *DHQ: Digital Humanities Quarterly*, *10*(3).

Social media platforms:
Who are interacting with each other?

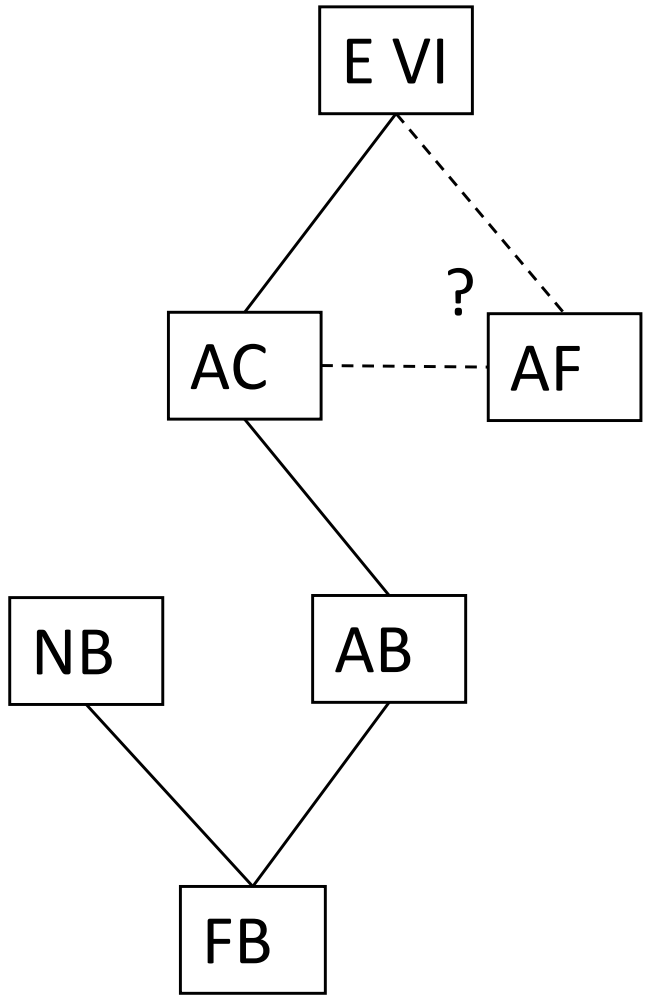
What if we want to learn an early social network?

What if the only source is nonstructural, historical text?

Learning from historical text

Biography Example

Bacon, Francis, Viscount St Alban (1561–1626), lord chancellor, politician, and philosopher, was born on 22 January 1561 at York House in the Strand, London, the second of the two sons of **Sir Nicholas Bacon** (1510–1579), lord keeper, and his second wife, **Anne** (c.1528–1610) [see **Bacon, Anne**], daughter of **Sir Anthony Cooke**, tutor to **Edward VI**, and his wife, **Anne, née Fitzwilliam**. He was baptized in the local church of St Martin-in-the-Fields, but spent most of his childhood, together with his elder brother, **Anthony Bacon** (1558–1601), at Gorhambury, near St Albans, Hertfordshire, which their father had purchased in 1557.

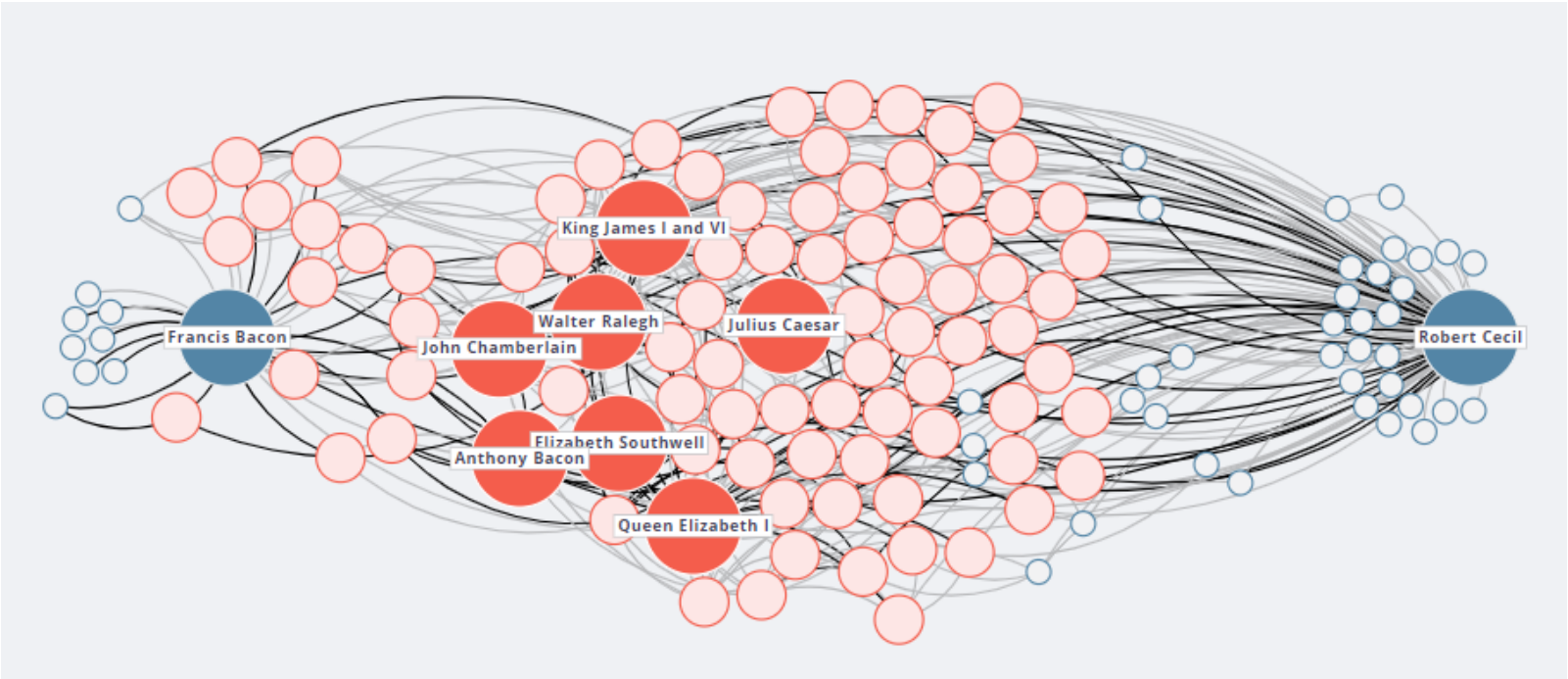


Previous work on recovering networks from text

- Match patterns to detect links
  - China Biographical Database Project (`projects.iq.harvard.edu/cbdb`)
- Co-occurrence to infer the network
  - Using radical environmentalist texts to uncover network structure and network features (Almquist & Bagozzi, 2019)
  - Mining and modeling character networks (Bonato et al., 2016)
- Need to manually clean the names
- Focus more on the analysis instead of the construction

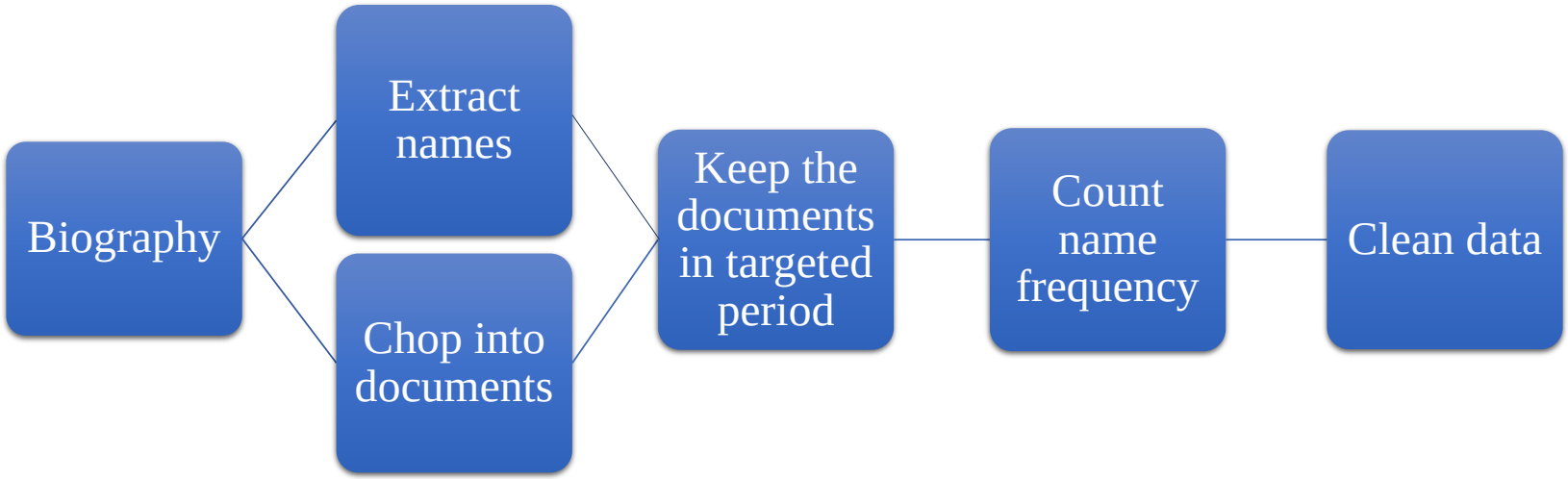
Introduction to the SDFB project

- Six Degree of Francis Bacon (SDFB) project (`https://www.sixdegreesoffrancisbacon.com/`)
- Digital reconstructions of an early modern social network (1500–1700) with Poisson Graphical Model
- The network is undirected and ignores the weight of edges



SDFB data and process

- **Data Source:** Oxford Dictionary of National Biography (ODNB) (`https://www.oxforddnb.com/`)
- Use Named Entity Recognizer (NER) to extract the names



- **Data Size:** 13309 name entities across 12149 biographies which result in 19686 documents

Example and Assumptions

Assumptions:

- If two people knew each other, it is more likely that they show up in the same biography or the same part of the biography (document).
- If two people are co-mentioned in the same document, it does not necessarily mean that they knew each other.

Goal: Apply conditional independence structure to illustrate relation

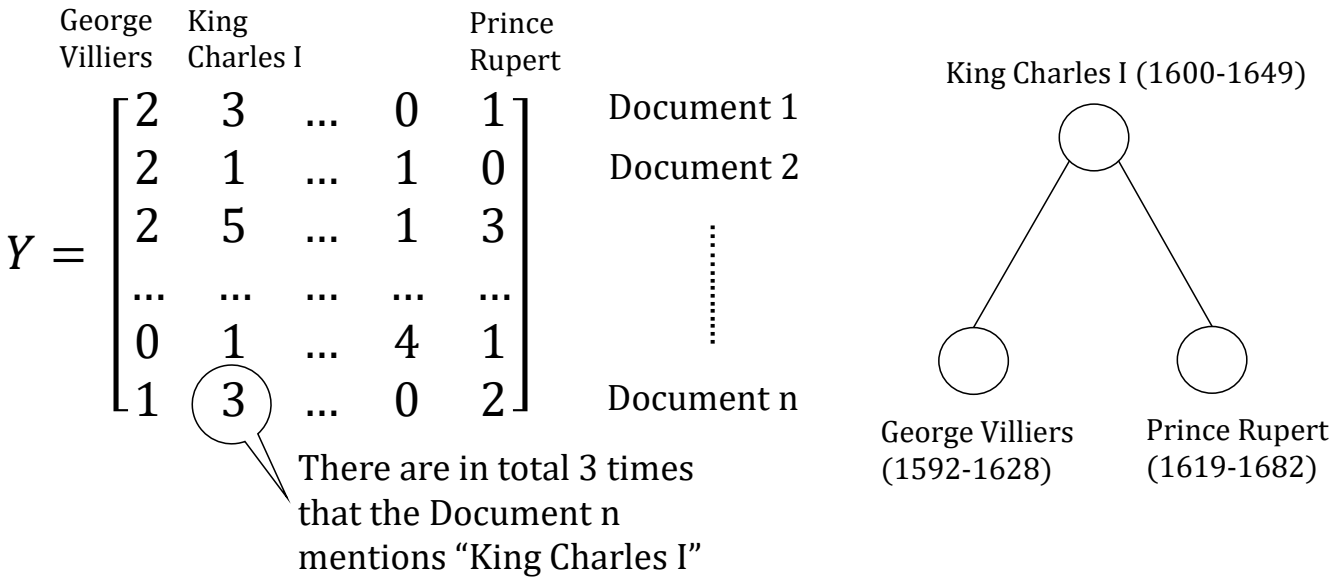


Figure: Data example (Warren et al., 2016)

Missing from SDFB

- Named Entity Recognizer (NER) does not always perfectly function
- Partial names and random names are not clear distinguished

Biography Example

[**Bacon**], [**Francis**], Viscount St Alban (1561–1626), lord chancellor, politician, and philosopher, was born on 22 January 1561 at York House in the Strand, London, the second of the two sons of Sir [**Nicholas Bacon**] (1510–1579), lord keeper, and his second wife, [**Anne**] (c.1528–1610) [see [**Bacon**],[**Anne**]], daughter of Sir [**Anthony Cooke**], tutor to [**Edward VI**], and his wife, [**Anne**], née Fitzwilliam. He was baptized in the local church of St Martin-in-the-Fields, but spent most of his childhood, together with his elder brother, [**Anthony Bacon**] (1558–1601), at Gorhambury, near St Albans, Hertfordshire, which their father had purchased in 1557.

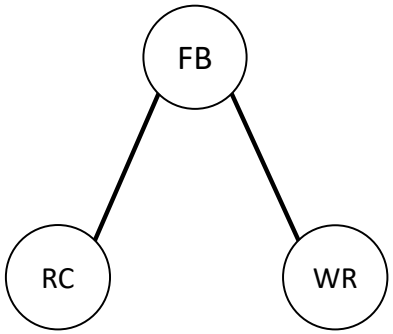
Name	Francis Bacon	Anne Bacon	Nicolas Bacon	Anthony Cooke
Count	?	?	?	1
Name	Edward VI	Anne Fitzwilliam	Anthony Bacon	
Count	1	?	?	

Missing from SDFB

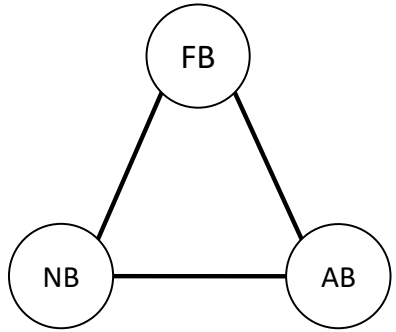
- Other available text information is not included
- Conditional independence structures $\neq$ social networks

Consider following two sentences:

Francis Bacon, like his friends Robert Cecil and Walter Raleigh, was a famous courtier.



Francis Bacon was the second son of Sir Nicolas Bacon and Anne Bacon.



Same count of names. Different network structure. Context is important.

Whose links are we missing?

- People who share the same characteristics should be more likely to be connected than those who are not:
  - People in the same family
  - People belonging to the same social group/club
  - People who work for the same company/serve the same king during the same period
  - People who live close to each other
- What if we make it easier for these people to be linked?

General Goal

General Goal: Improving the methodology of recovering social networks from the text data.

- Extend the current model to include covariates
- Incorporate record linkage techniques to assign the name mentions

Note: Today's talk will focus on model extension part.

Model in SDFB

- Local Poisson Graphical Lasso (Allen & Liu, 2012)
- Assumes that conditional on all the other people's mentions, the number of mentions for a person is a Poisson random variable

Model Definition

Suppose we only consider one document, and let Y_j denote how often person j is mentioned in this document and $Y_{\setminus j}$ contains the counts for all the other persons excluding j, then the model can be expressed as:

$$P(Y_j|Y_{\setminus j} = y_{\setminus j}, \theta, \Theta) \sim Poisson(\exp^{\theta_j + \sum_{k \neq j} \Theta_{jk} y_k})$$

- We interpret $\Theta_{jk} \leq 0$ as no connection between persons j, k.

Estimation of Local Poisson Graphical Lasso

- The probability density of the Poisson Markov Network is

$$P(Y|\theta,\Theta) = \exp\left\{\sum_j(\theta_j Y_j - \log(Y_j!)) + \sum_{j,k}\Theta_{jk}Y_j Y_k - A(\Theta)\right\}$$

  If we estimate globally all Θ_{jk} will be non-positive.

- Instead, estimate locally: for each person j, we fit a GLM and include an L1 penalty.
- A link between person j and person k exists if $\Theta_{jk} > 0$ or $\Theta_{kj} > 0$.
- The text information can be incorporated with the L1 penalty

Penalty factor matrix

- Background: We assume that two people's lifespan has to overlap for them to have been able to know each other
- SDFB solution: Remove non-overlap links post-modeling
- Our solution: Include in the model

Penalty factor matrix

We define the penalty factor matrix α, so that

$$\alpha_{jk} = \begin{cases} \infty & \text{If the lifespan of person } j \text{ and person } k \text{ did not overlap} \\ 1 & \text{Otherwise} \end{cases}$$

For a person j, our goal is to maximize the following penalized log-linear regression function:

$$\frac{1}{n}\sum_{i=1}^{n}[Y_{ij}(Y_{i,\neq j}\Theta_{\neq j,j}) - \exp(Y_{i,\neq j}\Theta_{\neq j,j})] - \rho||\alpha_{\neq j,j} \odot \Theta_{\neq j,j}||_1$$

Incorporate the covariates

The penalty factor matrix can be designed as a piece-wise function. Here is one heuristic example to include the covariate "last name" without training the parameter:

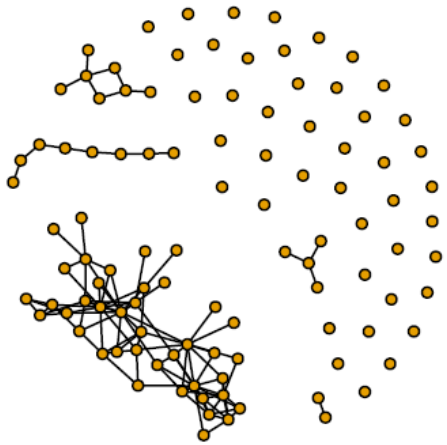
Penalty factor matrix

Let the penalty factor matrix α to be

$$\alpha_{jk} = \begin{cases} 0.5 & \text{person } j \text{ and person } k \text{ share the same last name and life span overlapped} \\ \infty & \text{person } j \text{ and person } k \text{ do not have life span overlapped} \\ 1 & \text{Otherwise} \end{cases}$$

SDFB Proof-of-Concept

Network with 104 people



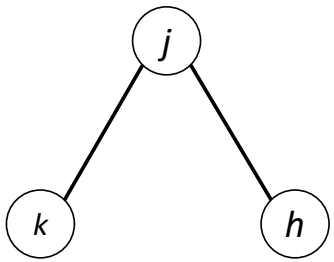
- 104 people with the most common 13 names, 800 documents, 107 links
- If two people's years do not overlap, set $\alpha = \infty$;
  If two people shared the same last name, set $\alpha = 0.5$
  - The current simulation method is based on Allen & Liu (2012)
    - Assume an adjacency matrix $A_{p\times p}$ is given
    - In the simulated data $Y_{n\times p}$, each observation Y_{ij} is sampled as

$$Y_{ij} \sim \sum_{m=1}^{p} A_{mj} Pois_{mj}(\mu) + Pois_{ij}(\epsilon)$$

where ϵ is the noise rate and μ is the true signal rate.

Simulation method example

Suppose we need to generate n documents for 3 people j, k and h and their network and corresponding adjacency matrix are:



$$\begin{array}{c} \\ j \\ k \\ h \end{array} \begin{array}{c} \begin{array}{ccc} j & k & h \end{array} \\ \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{array}$$

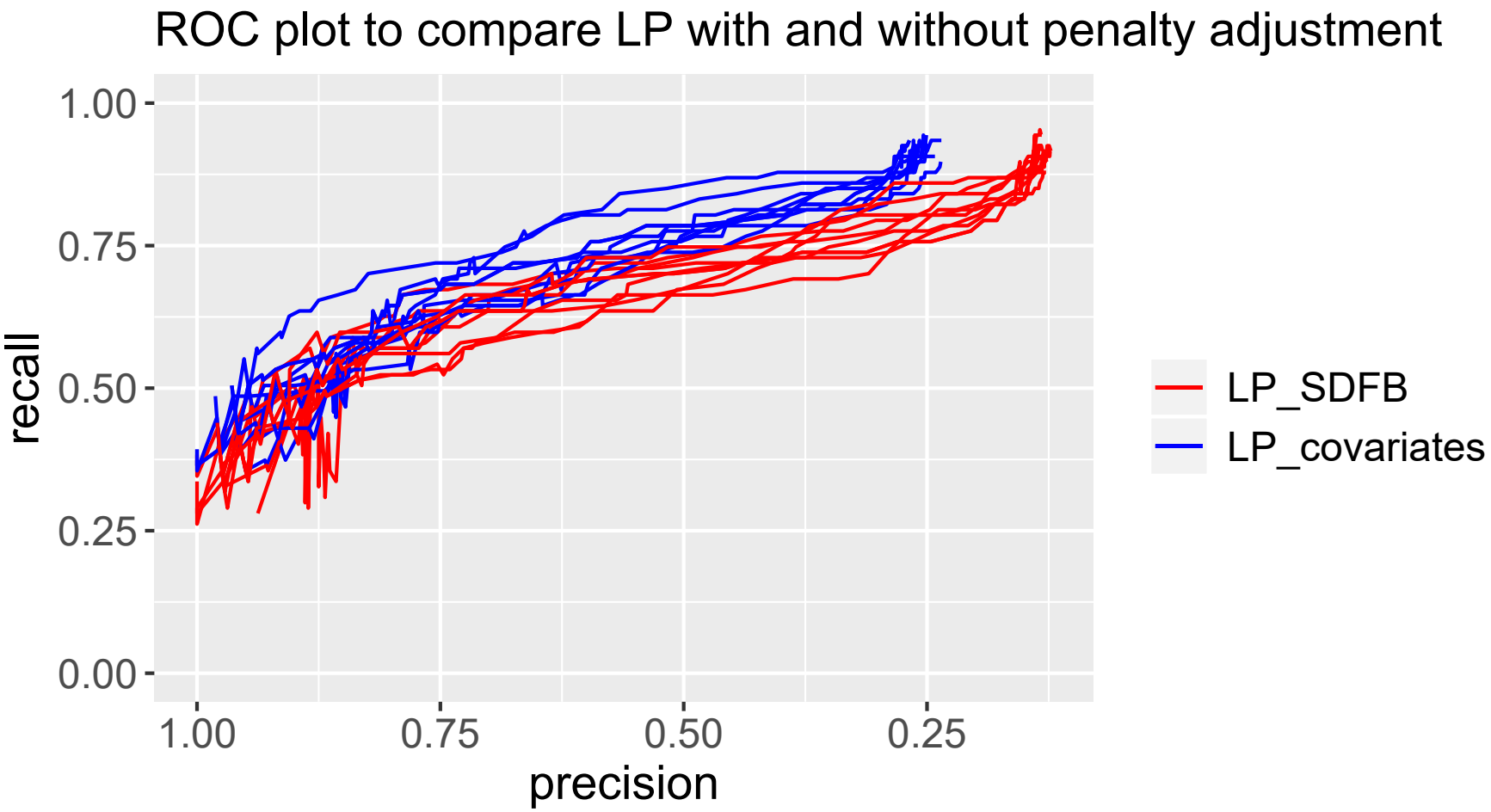
For one document: for each pair j and k, we generate a count $Poiss_{jk}(\mu)$, and for each person j, we generate a noise $Poiss_j(\epsilon)$. Then the total count for each person in the document will be:

$$\begin{array}{ll} \text{Person } j & \text{Poiss}_{jk}(\mu) + Poiss_{jh}(\mu) + Poiss_j(\epsilon) \\ \text{Person } k & \text{Poiss}_{jk}(\mu) + Poiss_k(\epsilon) \\ \text{Person } h & \text{Poiss}_{jh}(\mu) + Poiss_h(\epsilon) \end{array}$$

Repeat until you have n documents.

Simulated SDFB example

- Simulate 800 documents for 104 people, 107 possible links (10 times)
- Decrease computational complexity
- Reduce non-convergence issue



Advantages and next step

Advantages of incorporating multiple covariates into the model:

- Incorporate text information to better identify the social network; potentially benefit people less mentioned in documents
- can also include information of duplicated names as a model covariate eg: Three Francis Bacon in the same period,

Next step challenge:

- Incorporate any combination of covariates
- Quickly find the best penalty for each covariate

A more systematic way

Multiple penalty lasso

Suppose we have m covariates, and for each covariate, we define a connection matrix C^h such that

$$C^h_{jk} = \begin{cases} 1 & \text{person } j \text{ and person } k \text{ share the same characteristic } h \\ 0 & \text{Otherwise} \end{cases}$$

For a person j, our goal is to maximize the following penalized log-linear regression function:

$$\frac{1}{n}\sum_{i=1}^{n}[Y_{ij}(Y_{i,\neq j}\Theta_{\neq j,j}) - exp(Y_{i,\neq j}\Theta_{\neq j,j})] - \sum_{h=1}^{m}\rho_h||C^h_{\neq j,j} \odot \Theta_{\neq j,j}||_1$$

by training multiple ρ_h at the same time.

Next Steps: Incorporating Record Linkage

Biography Example

Bacon, Francis, Viscount St Alban (1561–1626), lord chancellor, politician, and philosopher, was born on 22 January 1561 at York House in the Strand, London, the second of the two sons of **Sir Nicholas Bacon** (1510–1579), lord keeper, and his second wife, **Anne** (c.1528–1610) [see **Bacon, Anne**], daughter of **Sir Anthony Cooke**, tutor to **Edward VI**, and his wife, **Anne, née Fitzwilliam**. He was baptized in the local church of St Martin-in-the-Fields, but spent most of his childhood, together with his elder brother, **Anthony Bacon** (1558–1601), at Gorhambury, near St Albans, Hertfordshire, which their father had purchased in 1557.

- Inference of the network much relies on the accuracy of the person-by-document matrix
- Duplicated and partial names are ubiquitous in historical data
- Associated with highest manual labeling cost
- No ground truth to evaluate the assignments

Record Linkage in the name assignment

Our solution:

- Incorporate covariates and apply record linkage techniques
- The count will be assigned to the people who are most **similar** to the biography owner

Document: King Charles I (1600-1649)

A name mentioned : Francis Bacon

Name	Sir Francis Bacon	Francis Bacon, MP	Sir Francis Bacon
Birth/death year	1561-1626	1600-1663	1587-1657
Biography Length	**22100**	**632**	**606**
Occupation	lord chancellor, politician, and philosopher	politician	judge

Takeaway

- Text data can contain useful information about the social network.
- Need systematic way to extract information about both the people and the network from the text
- Including context covariates can better identify the social network
- Recording Linkage technique can be used to identify the duplicated names in historical text

Contact Information:

- Email: xiaoyiy@andrew.cmu.edu
- Webpage: https://sites.google.com/view/xiaoyiyang

Thank you and questions?

Reference I

John Aitchison and CH Ho.

The multivariate poisson-log normal distribution.
Biometrika, 76(4):643–653, 1989.

Genevera I Allen and Zhandong Liu.

A log-linear graphical model for inferring genetic networks from high-throughput sequencing data.
pages 1–6, 2012.

Zack W Almquist and Benjamin E Bagozzi.

Using radical environmentalist texts to uncover network structure and network features.
Sociological Methods & Research, 48(4):905–960, 2019.

Anthony Bonato, David Ryan D'Angelo, Ethan R Elenberg, David F Gleich, and Yangyang Hou.

Mining and modeling character networks.
In *International workshop on algorithms and models for the web-graph*, pages 100–114. Springer, 2016.

Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin.

Variational inference for sparse network reconstruction from count data.
arXiv preprint arXiv:1806.03120, 2018.

Yoonha Choi, Marc Coram, Jie Peng, and Hua Tang.

A poisson log-normal model for constructing gene covariation network using rna-seq data.
Journal of Computational Biology, 24(7):721–731, 2017.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning.

Incorporating non-local information into information extraction systems by gibbs sampling.
In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

Reference II

Jerome Friedman, Trevor Hastie, and Robert Tibshirani.

Sparse inverse covariance estimation with the graphical lasso.
Biostatistics, 9(3):432–441, 2008.

Gene H Golub, Michael Heath, and Grace Wahba.

Generalized cross-validation as a method for choosing a good ridge parameter.
Technometrics, 21(2):215–223, 1979.

Academia Sinica Harvard University and Peking University.

China biographical database project (cbdb).
https://projects.iq.harvard.edu/cbdb.
Accessed: 2020-02-21.

Steffen L Lauritzen.

Graphical models, volume 17.
Clarendon Press, 1996.

Peter V Marsden.

Network data and measurement.
Annual review of sociology, 16(1):435–463, 1990.

Henry Colin Gray Matthew, Brian Harrison, and R James Long.

The Oxford dictionary of national biography.
R. James Long (2004)., 2004.

David Sinclair and Giles Hooker.

Sparse inverse covariance estimation for high-throughput microrna sequencing data in the poisson log-normal graphical model.
Journal of Statistical Computation and Simulation, 89(16):3105–3117, 2019.

Reference III

Behlül Üsdiken and Yorgo Pasadeos.

Organizational analysis in north america and europe: A comparison of co-citation networks.
Organization studies, 16(3):503–526, 1995.

Christopher N Warren, Daniel Shore, Jessica Otis, Lawrence Wang, Mike Finegold, and Cosma Shalizi.

Six degrees of francis bacon: A statistical method for reconstructing large historical social networks.
DHQ: Digital Humanities Quarterly, 10(3), 2016.

Estimation of Poisson Graphical Lasso

- The model is defined locally and if we want all local, pair-wise Markov property to hold, then the probability density of this the Poisson Markov Network is

$$P(Y|\theta,\Theta) = exp\{\sum_j(\theta_j Y_j - log(Y_j!)) + \sum_{j,k}\Theta_{jk} Y_j Y_k - A(\Theta)\}$$

- Y is count matrix, so the range of value is from 0 to ∞. Also, $log(Y_j!)$ decreases much slower than $\Theta_{jk} Y_j Y_k$. If any Θ_{jk} is positive, we cannot find a finite $A(\Theta)$ so that $P(Y|\theta,\Theta)$ is a probability.
- However, a negative Θ_{jk} is not meaningful in the context, which suggests that the count of people j will decrease if we mention people k more in the document.

Estimation of Gaussian Graphical Lasso

- The estimation of Ω is though `gLASSO` package in R (Friedman, Hastie & Tibshirani, 2008), which mximize the L1 penalized log-likelihood

$$\log\det\Omega - trace(S\Omega) - \lambda||\Omega||_1$$

with corrdinate descent, and S is the empirical covariance matrix Instead of directly using the count matrix as the input, the model takes the empirical covariance matrix S to infer the precision matrix Ω, which represents the partial correlations.

- The estimation is global

- The estimation starts with the empirical covariance matrix S, and then globally estimating one row and one column at the same time until the results converge, so that $\Omega_{jk} = \Omega_{kj}$ in the final Ω.

- Need to choose a penalty parameter, either through cross-validation or stability selection

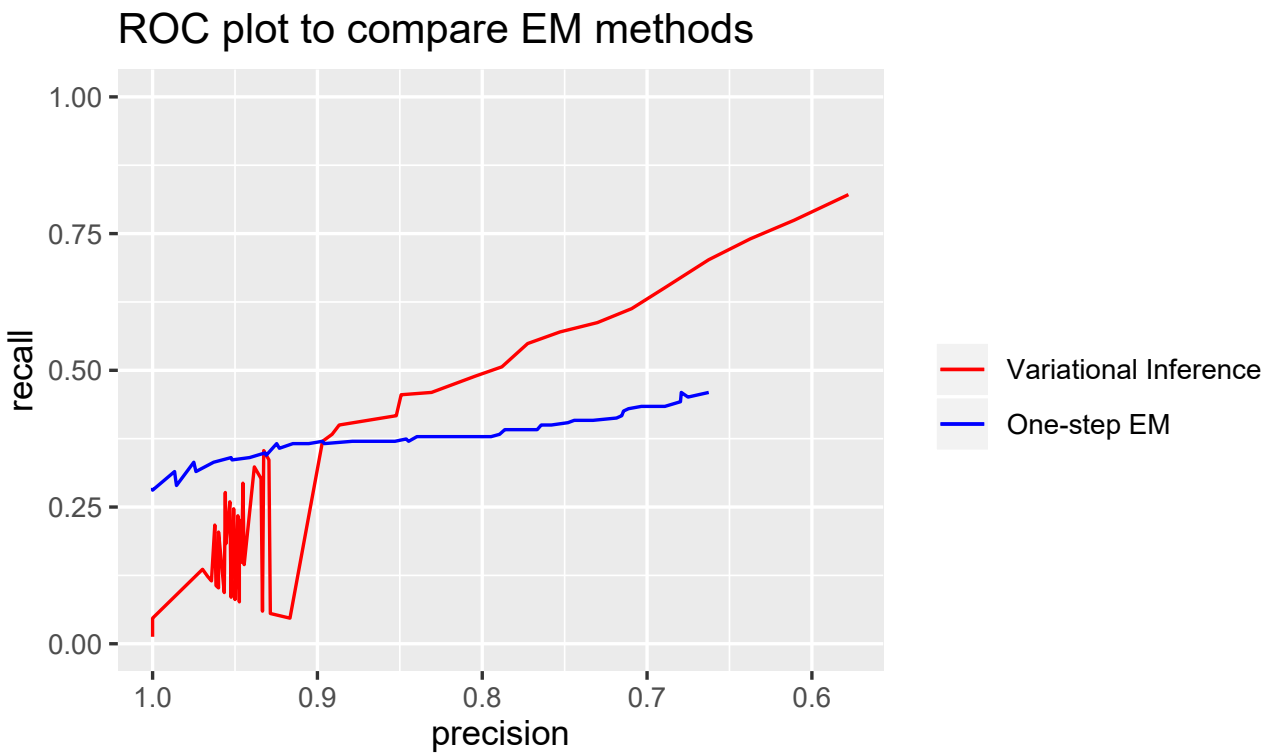
Estimation of Poisson Log Normal

- Latent variable is unobserved, cannot evaluate the log likelihood of the observed data $\log p_\Omega(Y) = \log \int p_\Omega(Y, Z) dZ$
- *Solution 1*: Approximate the l_1 penalized log-likelihood function with Laplace's method of integration (Choi et al, 2017)
- *Solution 2* : EM algorithm
  - Initiate Σ as a diagonal matrix and transform each data to the posterior mean $E_{Z_{ij}|Y_{ij}}(Z_{ij})$ with one step EM algorithm. (Sinclair & Hooker, 2019)
  - Variational inference, approximate conditional distribution $p_\Omega(Z_i|Y_i)$ with a set of multivariate Gaussian distributions. (Chiquet, Mariadassou & Robin, 2018)

We use the last one for computational reasons.

Comparison on PLN estimation

- Laplace's method is much slower than EM algorithm
  (1 hour vs 8 minutes for 100 cross-validation)
  Laplace's via R package, need to choose Lagrange parameter
- Implemented and adjusted one-step EM to be feasible for sparse data



Review on Generalized cross-validation

- In Ridge regression, for penalty parameter λ

$$\hat{\beta}(\lambda) = (X^T X + n\lambda I)^{-1} X^T Y$$

- Consider a $n \times n$ influence matrix $A(\lambda)$, and if we let

$$A(\lambda) = X(X^T X + n\lambda I)^{-1} X^T$$

then we will have

$$\hat{Y} = X\hat{\beta} = A(\lambda) Y$$

- In this case, we can let $A(\lambda)$ serves as a "hat matrix"
- Then the GCV estimate of λ in the ridge estimate can be written as

$$V(\lambda) = \frac{\frac{1}{n}||(I - A(\lambda))Y||^2}{\frac{1}{n} tr(I - A(\lambda))}$$

Record Linkage in the name assignment

- Utilize the text information to identify the most probable candidate when a name is mentioned
- Assumption: When a name is mentioned in person j's biography, the count should go to the person with the same name and also who is "most similar" candidate to person j. Only one candidate is selected.
- Current available variables:
  - Birth and death year: can used to calculate people's overlapping years
  - Each candidate's biography length: as a signal for the popularity
- No model-based methods have been considered so far because of the limited number of covariates
- Instead, calculate a similarity score, can extend to similarity fucntions later

Similarity Score

The biography is about person k and a name is mentioned. For each candidate person k with that mentioned name, we can define the similarity level s_k as

$$s_j = \frac{\text{intersect lifespan between } j \text{ and } k}{\text{union lifespan between } j \text{ and } k} \times \frac{\log(\text{person } j\text{'s biography length})}{\log(\text{max biography length})},$$

if person j do not have a biography, the later part will be set as 0.5. The distribution of biography length is likely to be high-skewed so a logarithm is included (may not necessary). Then the probability that the count will assign to person j is

$$p(\text{person } k \text{ got the count}) = \frac{s_k}{\sum_m s_m}.$$

We can either randomly assign the count based on the probability, or simply assign the count to the candidate with the highest probability.

Certainty of linkage as a covariate

Consider a link between person j and k, and there are m people with the same name as person j and n people with the same name as person k:

- If person j and person k have different names
  - They should be less likely to be linked if duplicated name exists
  - An easy approach: create a connection matrix C^1 and we define

$$C^1_{jk} = \begin{cases} mn & \text{person } j \text{ and person } k \text{ have different names and } mn > 1 \\ 0 & \text{Otherwise} \end{cases}$$

    and a penalty is trained for this case.

- If person j and person k have the same names
  - This is a mix situation when people with same name can be either strangers or families, may want to use birth/death year to distinguish.
  - Create a separate connection matrix C^2 and we define

$$C^2_{jk} = \begin{cases} 0 & \text{Otherwise} \\ (m-1)^2 & \text{person } j \text{ and person } k \text{ have the same name} \end{cases}$$

    and another penalty is trained for this case.