A Bayesian Idealization of Entity Resolution on Networks

Jim Ferry, Darren Lo, and Thomas Seaquist
Metron, Inc.

Graph Exploitation Symposium
Wednesday, May 18, 2016
MIT Endicott House

METRON
SCIENTIFIC SOLUTIONS

Overview

- Motivation
  - General: develop idealized models of practical problems
  - Specific: the Bhattacharya–Getoor algorithm for entity resolution
- Elements of the model
  - Keep everything as simple as possible, but no simpler
  - Want equivalent of Gaussian, Poisson, or Erdős–Rényi distribution
  - Example: generate a sample
- Derivation of entity resolution probabilities
  - Application to simple example
- Comparison to algorithm

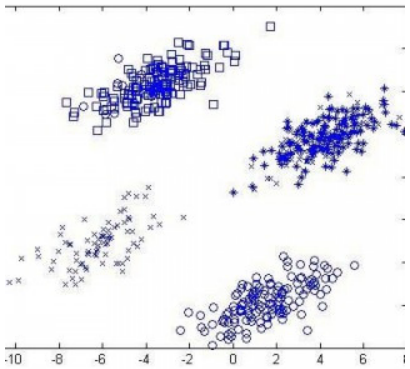
General Motivation

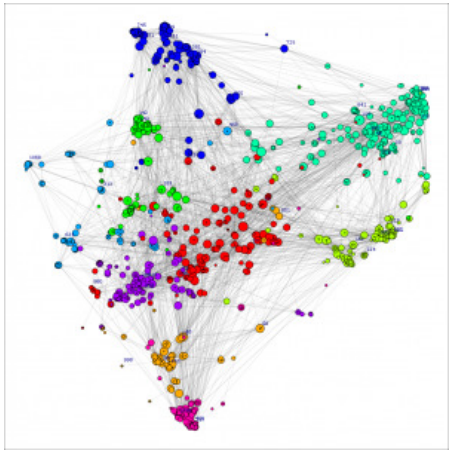
- Idealize real-world problems
  - To create interesting mathematical structures
  - In the hopes of motivating powerful new abstractions
- Algorithms vs. generative models
  - Algorithms
    - Easier to develop
    - Have more practical use
    - Work well when a single correct answer exists
    - Focus on efficiency for big data
  - Generative models
    - Lead to deeper understanding
    - Support simulation
    - Yield posterior distributions over many possible answers
    - Can be applied to (though sometimes limited to) small data

METRON
SCIENTIFIC SOLUTIONS

Specific Motivation

- The Bhattacharya–Getoor algorithm
  - Data: network with noisy node labels (*aliases*)
    - E.g., multigraph of *trades* between companies
    - But company names are poorly recorded
  - Goal: correctly partition aliases into entities
  - Intuition: grouping two aliases favored when
    - Nearby in alias space (e.g., string metric)
    - Their trading partners are the same entities
  - Method:
    - Combine alias and network similarities
    - Iterate with better partitions for network similarity
- Specific motivation
  - Generative model for Bhattacharya–Getoor scenario
  - E.g., what is probability of any partition hypothesis?

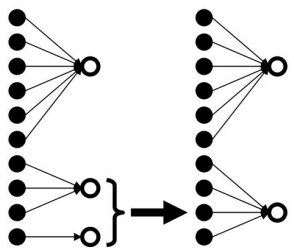




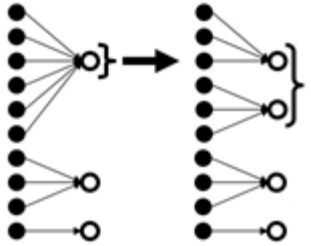
METRON
SCIENTIFIC SOLUTIONS

Entity Resolution in Transaction Networks

- Let's suppose
  - *Entities* exist in some *state space*
  - Entities engage in *transactions*, but
  - Only the entities' *aliases* are observed in the transactions
- Imagine the state space as a *continuum*
  - Want to avoid messy complications in general
  - Want to avoid accidental *conjunctive ambiguities*
    - Zero probability of picking same point twice "by accident"
- Disjunctive ambiguity simpler
  - Merge groups? Fine.
- Conjunctive harder
  - Split this group?
  - Who goes where?



Resolving a disjunctive ambiguity

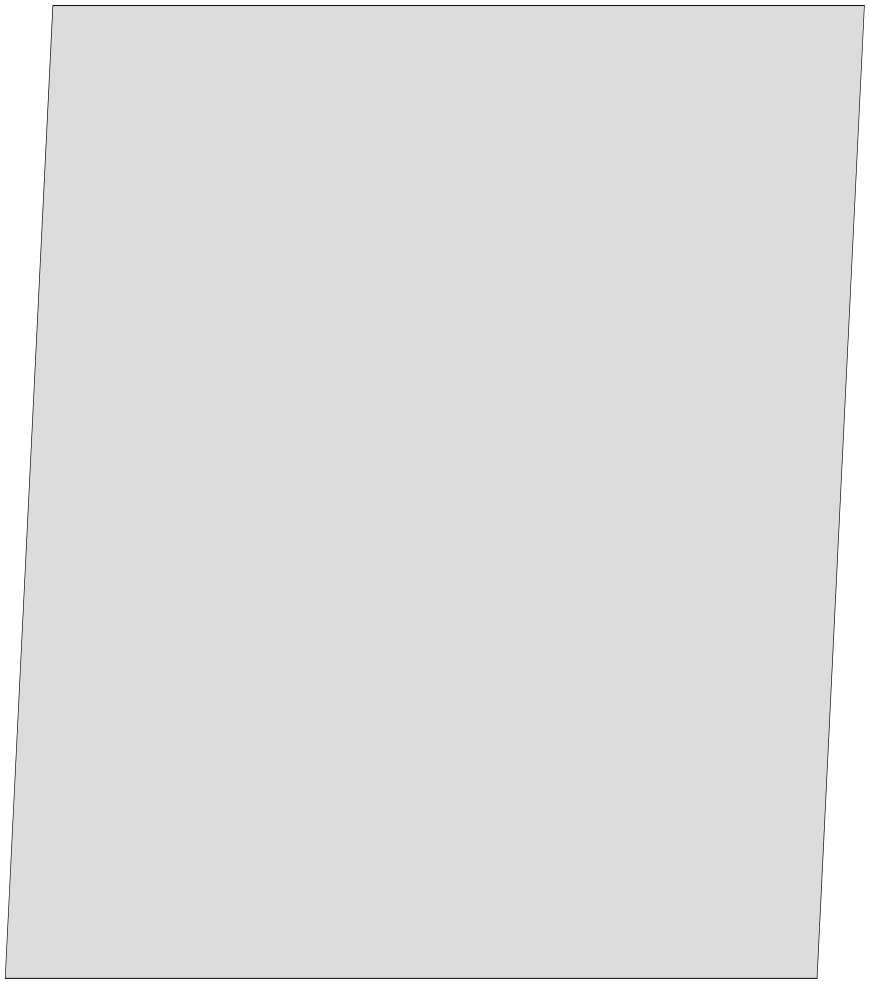


Resolving a conjunctive ambiguity

METRON SCIENTIFIC SOLUTIONS

Okay! Let's Model!

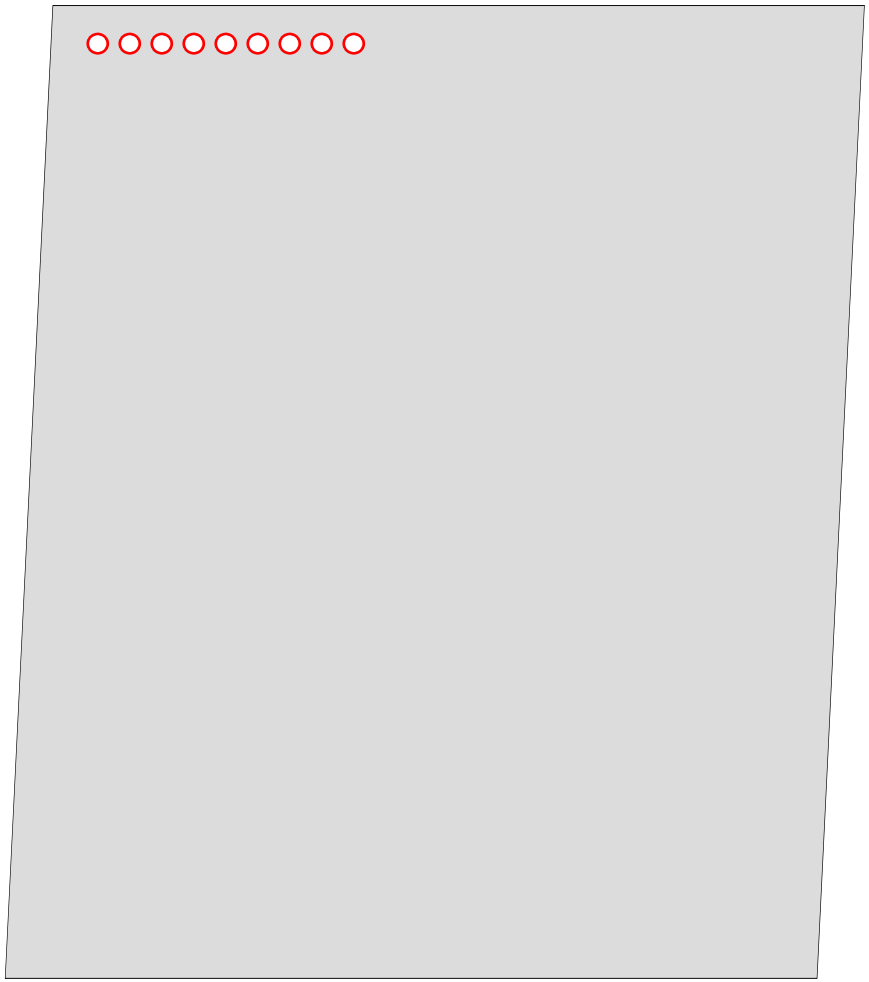
- Causal model for situation
  - Prior on number of entities: $\rho^0(n)$
  - Prior on state: $p^0(x)$
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - Likelihoods for aliases: $L(z|x)$
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\Pr([a]|Z)$
- Result…



METRON
SCIENTIFIC SOLUTIONS

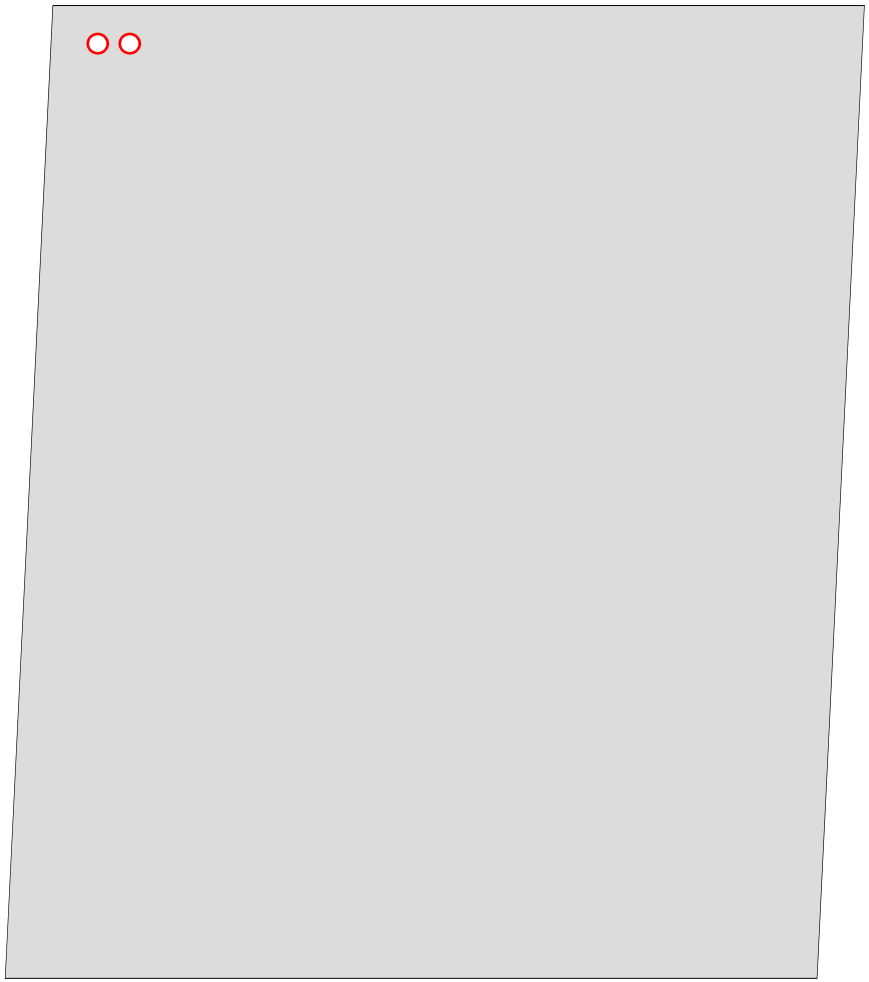
Okay! Let's Model!

- Causal model for situation
  - **Prior on number of entities: $\rho^0(n)$**
  - Prior on state: $p^0(x)$
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - Likelihoods for aliases: $L(z|x)$
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\Pr([a]|Z)$
- Result…



Okay! Let's Model!

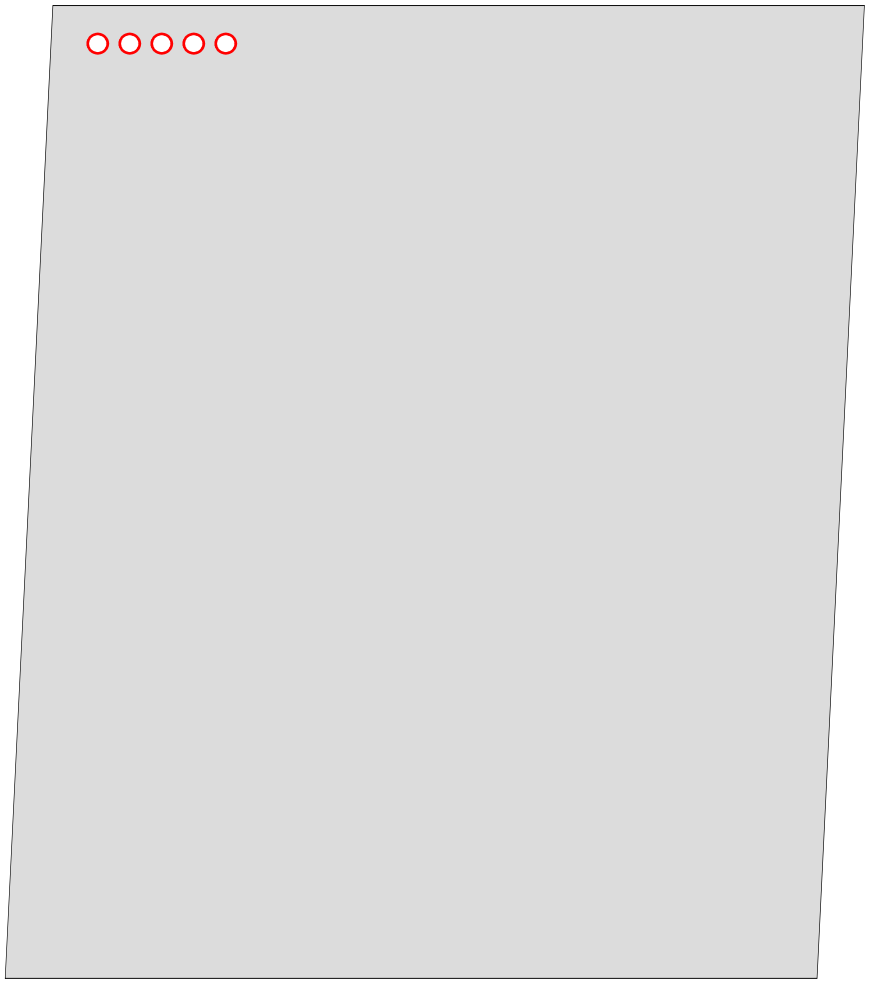
- Causal model for situation
  - **Prior on number of entities: $\rho^0(n)$**
  - Prior on state: $p^0(x)$
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - Likelihoods for aliases: $L(z|x)$
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\Pr([a]|Z)$
- Result…



METRON
SCIENTIFIC SOLUTIONS

Okay! Let's Model!

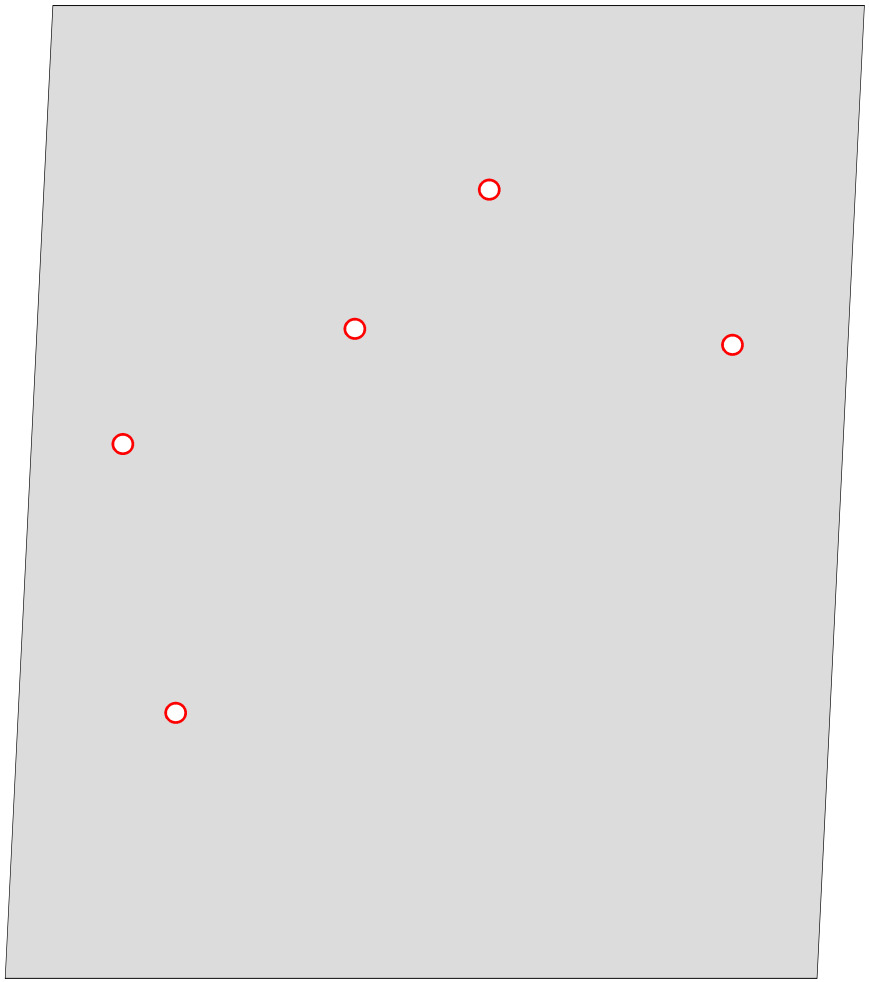
- Causal model for situation
  - **Prior on number of entities: $\rho^0(n)$**
  - Prior on state: $p^0(x)$
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - Likelihoods for aliases: $L(z|x)$
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\Pr([a]|Z)$
- Result…



METRON SCIENTIFIC SOLUTIONS

Okay! Let's Model!

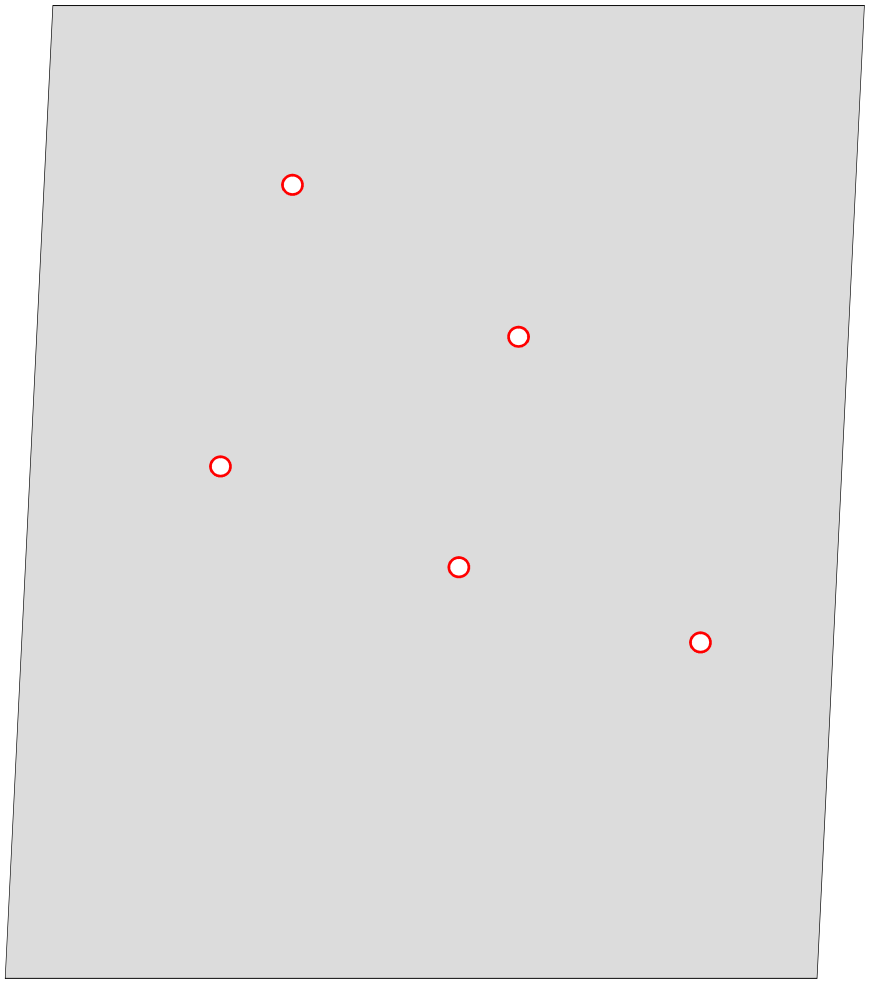
- Causal model for situation
  - Prior on number of entities: $\rho^0(n)$
  - **Prior on state: $p^0(x)$**
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - Likelihoods for aliases: $L(z|x)$
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\Pr([a]|Z)$
- Result…



METRON
SCIENTIFIC SOLUTIONS

Okay! Let's Model!

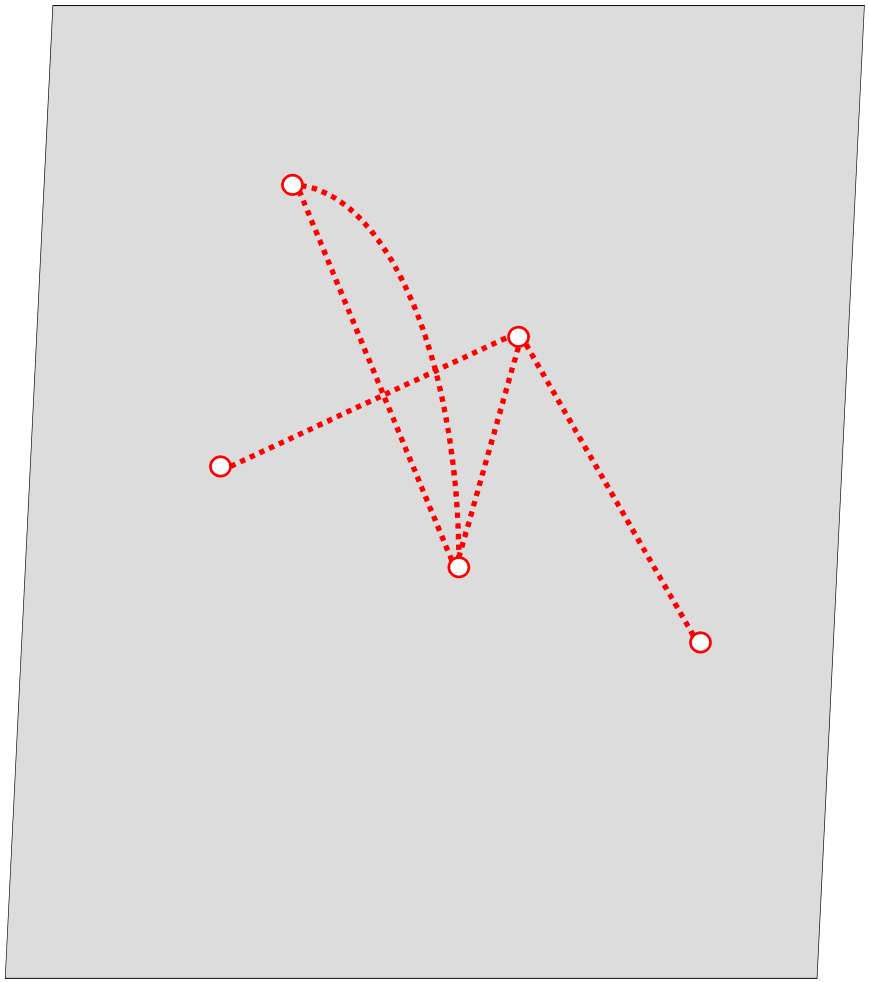
- Causal model for situation
  - Prior on number of entities: $\rho^0(n)$
  - **Prior on state: $p^0(x)$**
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - Likelihoods for aliases: $L(z|x)$
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\Pr([a]|Z)$
- Result…



METRON
SCIENTIFIC SOLUTIONS

Okay! Let's Model!

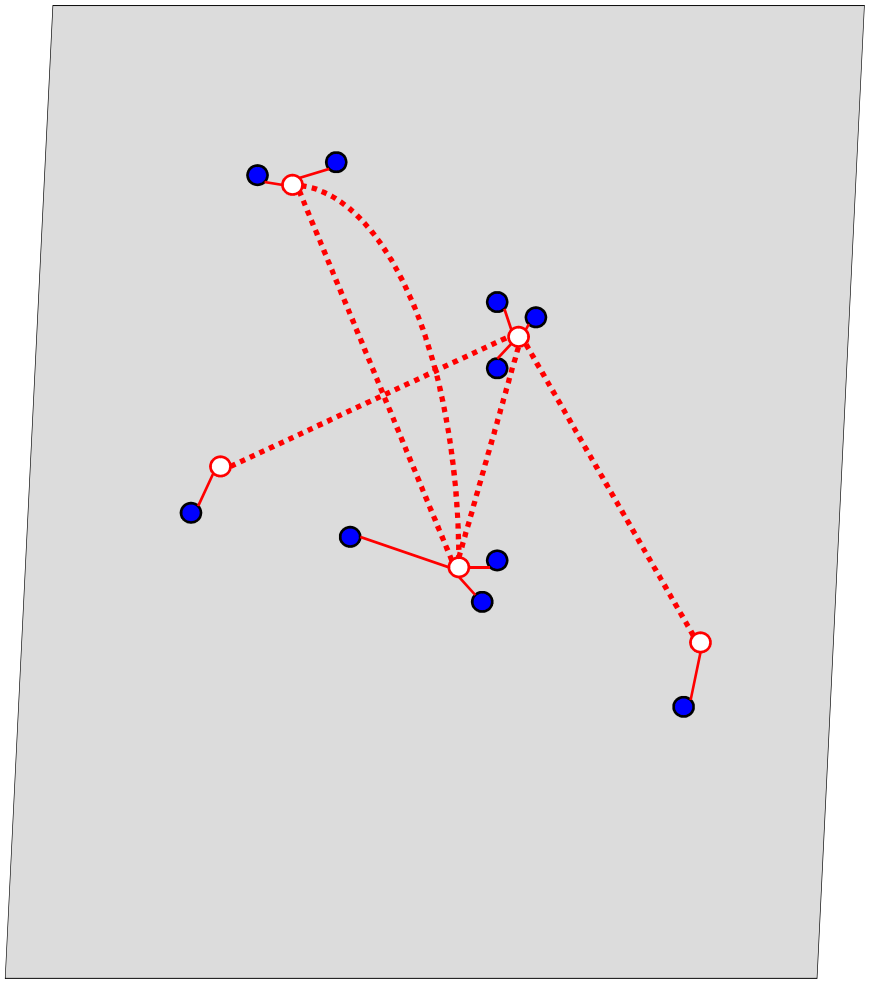
- Causal model for situation
  - Prior on number of entities: $\rho^0(n)$
  - Prior on state: $p^0(x)$
  - **Transaction rate: $\mathbf{Po}(\lambda t)$**
  - Likelihoods for aliases: $L(z|x)$
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\Pr([a]|Z)$
- Result…



METRON
SCIENTIFIC SOLUTIONS

Okay! Let's Model!

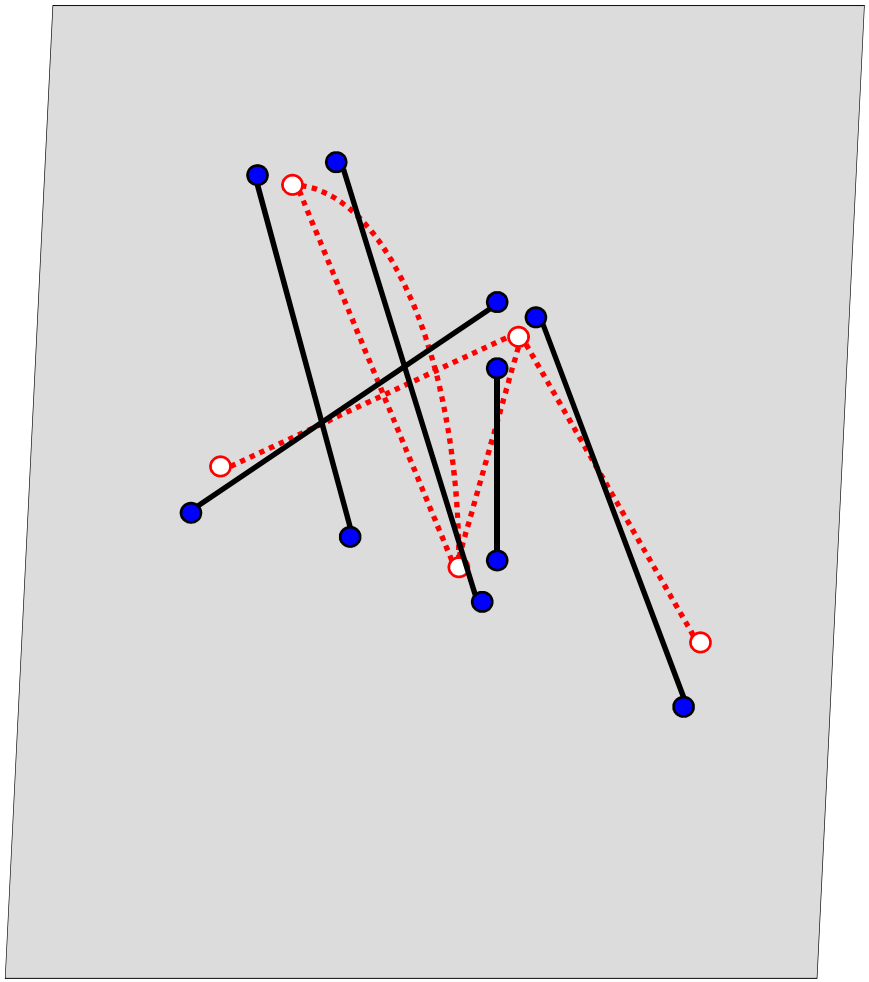
- Causal model for situation
  - Prior on number of entities: $\rho^0(n)$
  - Prior on state: $p^0(x)$
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - **Likelihoods for aliases: $L(z|x)$**
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\Pr([a]|Z)$
- Result…



METRON
SCIENTIFIC SOLUTIONS

Okay! Let's Model!

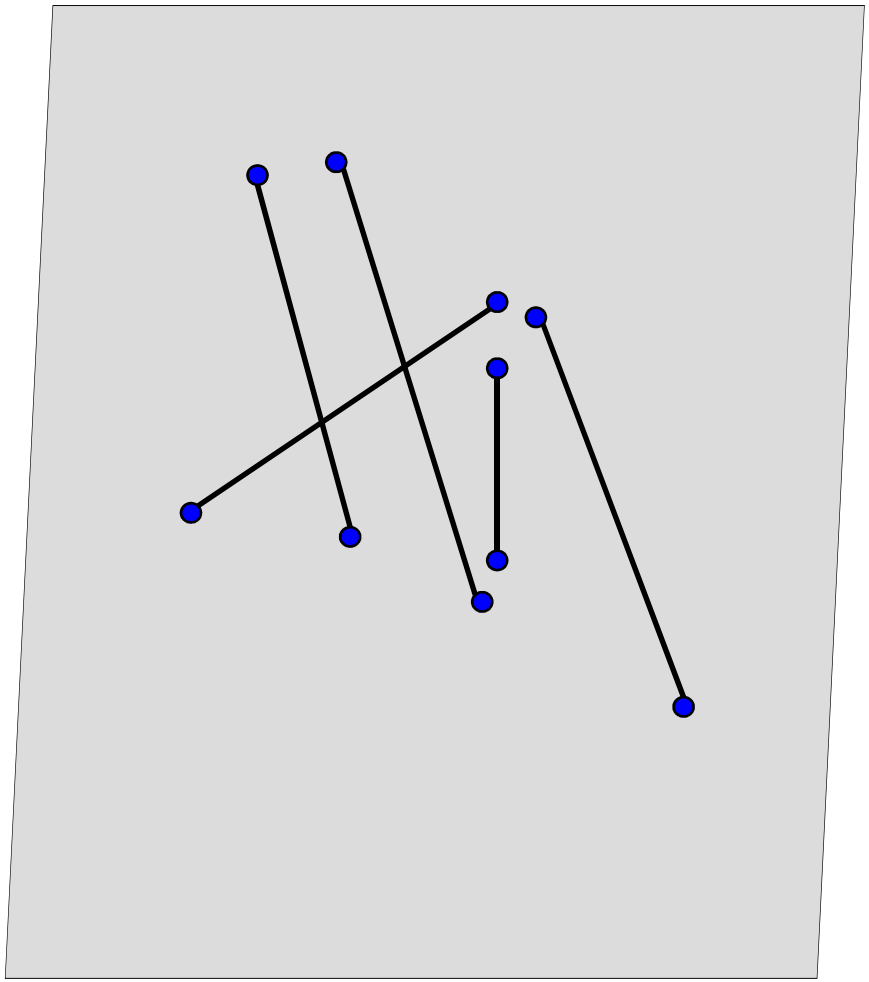
- Causal model for situation
  - Prior on number of entities: $\rho^0(n)$
  - Prior on state: $p^0(x)$
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - **Likelihoods for aliases: $L(z|x)$**
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\Pr([a]|Z)$
- Result…



METRON
SCIENTIFIC SOLUTIONS

Okay! Let's Model!

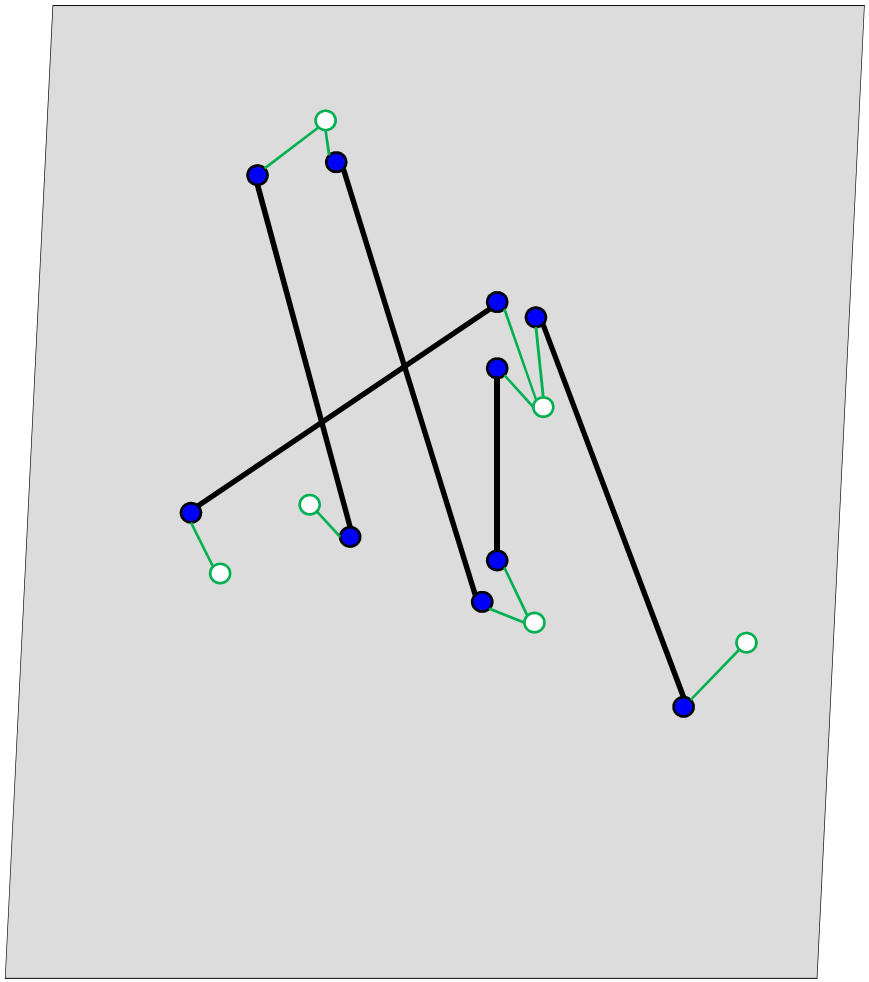
- Causal model for situation
  - Prior on number of entities: $\rho^0(n)$
  - Prior on state: $p^0(x)$
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - **Likelihoods for aliases: $L(z|x)$**
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\Pr([a]|Z)$
- Result…



METRON
SCIENTIFIC SOLUTIONS

Okay! Let's Model!

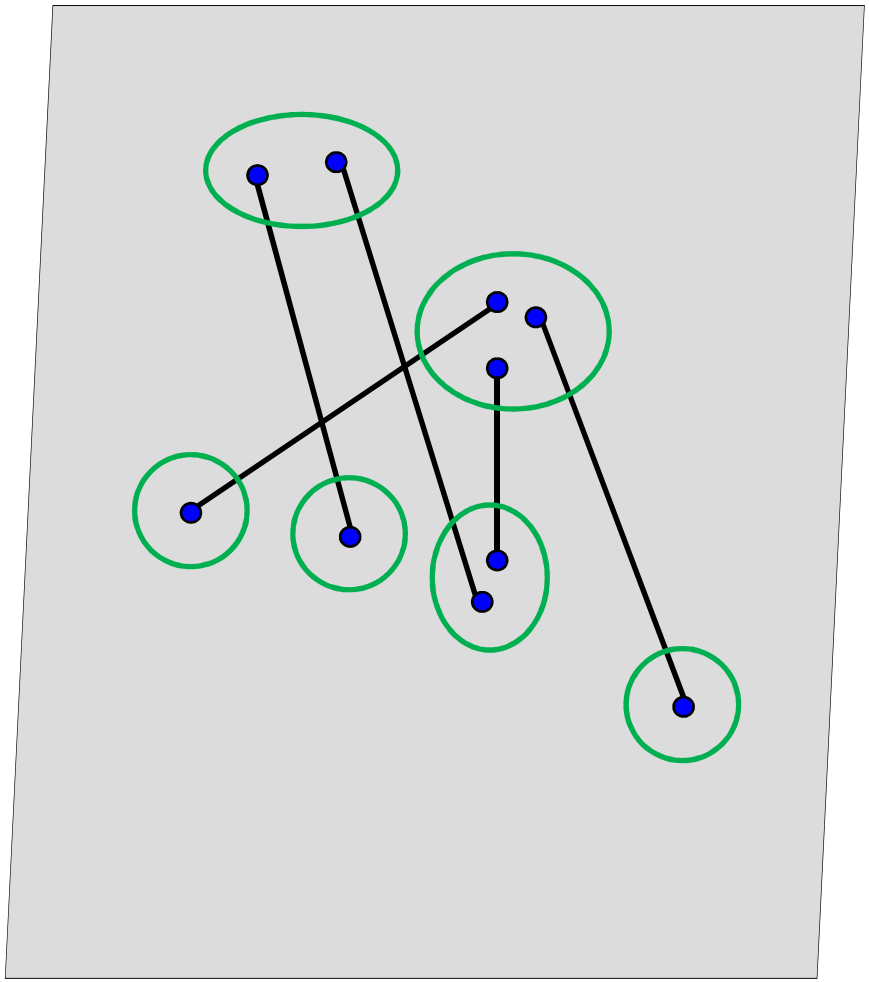
- Causal model for situation
  - Prior on number of entities: $\rho^0(n)$
  - Prior on state: $p^0(x)$
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - Likelihoods for aliases: $L(z|x)$
    - Perhaps Gaussians…
- Derive association probability
  - $\mathbf{Pr}(\boldsymbol{Z,a|x_1,\ldots,x_n,n})$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\mathrm{Pr}([a]|Z)$
- Result…



METRON
SCIENTIFIC SOLUTIONS

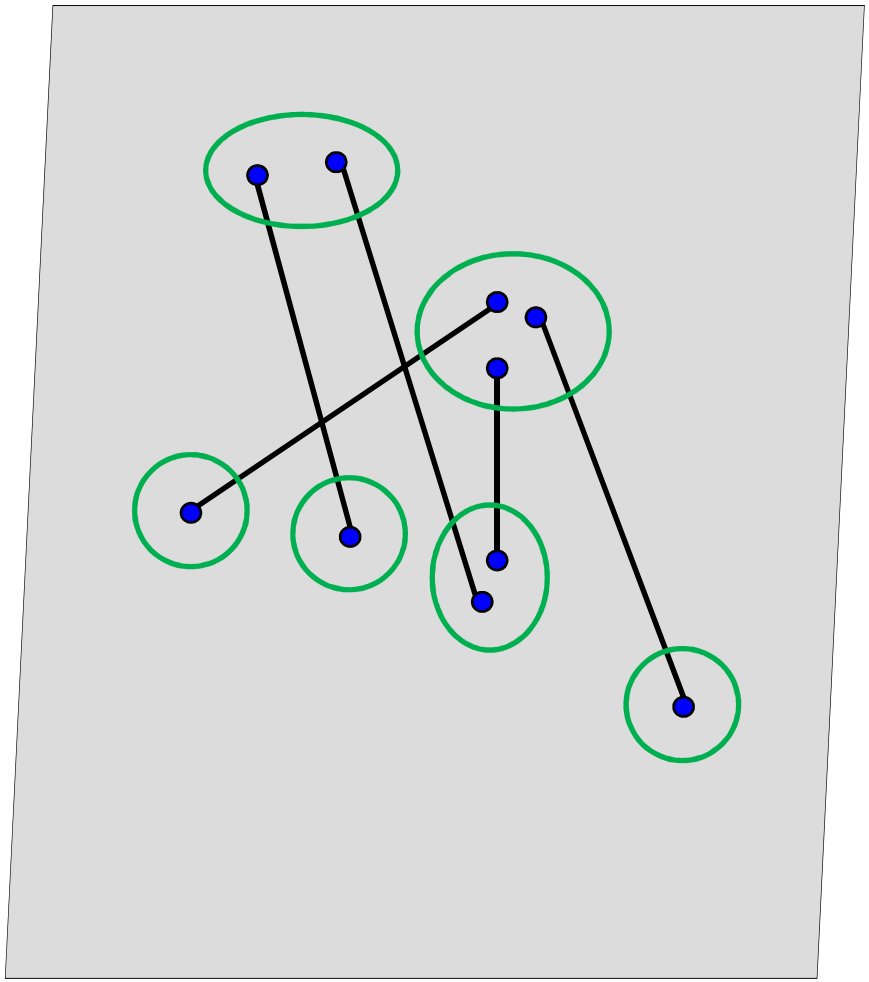
Okay! Let's Model!

- Causal model for situation
  - Prior on number of entities: $\rho^0(n)$
  - Prior on state: $p^0(x)$
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - Likelihoods for aliases: $L(z|x)$
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - **Bayesian inversion: $\Pr([a]|Z)$**
- Result…



Okay! Let's Model!

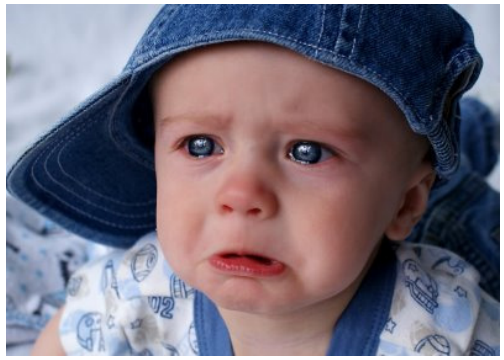
- Causal model for situation
  - Prior on number of entities: $\rho^0(n)$
  - Prior on state: $p^0(x)$
  - Transaction rate: $\mathrm{Po}(\lambda t)$
  - Likelihoods for aliases: $L(z|x)$
    - Perhaps Gaussians…
- Derive association probability
  - $\Pr(Z,a|x_1,\ldots,x_n,n)$
  - Integrate with $p^0(x_1)\cdots p^0(x_n)\rho^0(n)$
  - Bayesian inversion: $\Pr([a]|Z)$
- Result…
  - Terrible!
    - Models should be as simple as possible… but no simpler



METRON
SCIENTIFIC SOLUTIONS

What's Wrong? Two Things…

- Aliases should recur
  - Expect some common aliases, plus a long tail
  - Instead, every alias is unique
  - And the graph is just a "bag of links"
  - Identical aliases a key indicator of same entity



- Worse, links are totally non-informative
  - A and B both trade with C…
    - Should increase the odds that A and B refer to the same entity
    - Instead, all it means is "hey, the model has links"
- How to fix it?
  - Need model that makes aliases recur
  - Need an underlying network of trading partners
  - But still want to keep it simple and natural…

METRON
SCIENTIFIC SOLUTIONS

Simple and Natural

- Model for trading partner graph: Erdős–Rényi $\mathcal{G}(n,p)$
  - Simplest, most generic model, but enough to generate desired effect
- Model for alias generation
  - I.i.d. draws too simple: previous labels should arise again unstipulated
  - Relax to *exchangeable* process: order does not matter
  - (de Finetti–Hewitt–Savage) Any exchangeable process equivalent to
    - Drawing a distribution μ from some prior $\mathcal{D}$ (over all distributions)
    - Then drawing i.i.d. samples from μ
  - In finite setting (e.g., a finite set S of strings) this means
    - Drawing a probability vector $\mathbf{p} = (p_s)_{s \in S}$ from some prior distribution $\mathcal{D}$
    - Drawing string s with probability p_s
  - Dirichlet distribution $\mathcal{D} = \mathrm{Dir}(\alpha)$ is the natural (conjugate) prior for $\mathbf{p}$
  - Continuum limit is the Dirichlet process – equivalent to
    - Selecting cluster sizes from a Chinese Restaurant Process
    - Distributing clusters according to some base distribution (e.g., Gaussian)

METRON SCIENTIFIC SOLUTIONS

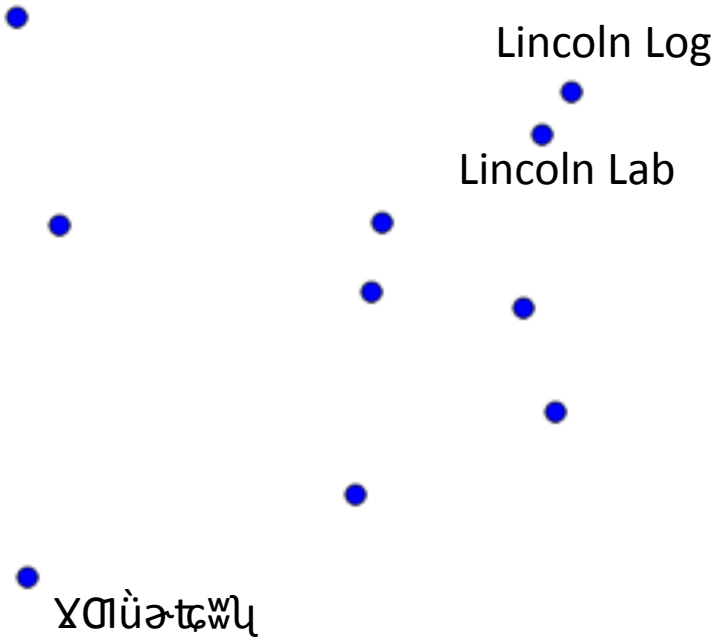
Let's Remodel…

- Causal model for situation
  - Prior on number of entities: $\rho^0(n)$
  - Prior on state: $p^0(x)$
  - Prior on *trading partners*: Erdős–Rényi random graph $\mathcal{G}(n,p)$
  - Transaction rate (for trading partners only): $\mathrm{Po}(\lambda t)$
  - Likelihoods for aliases: $L(\mathbf{z}|x)$
    - Not a Gaussian for individual aliases z
    - But a Dirichlet Process for a collection of aliases $\mathbf{z}$
      - With a Gaussian base distribution,
      - And a concentration parameter θ for the clustering
- It's a generative model, so we can generate a sample

METRON SCIENTIFIC SOLUTIONS

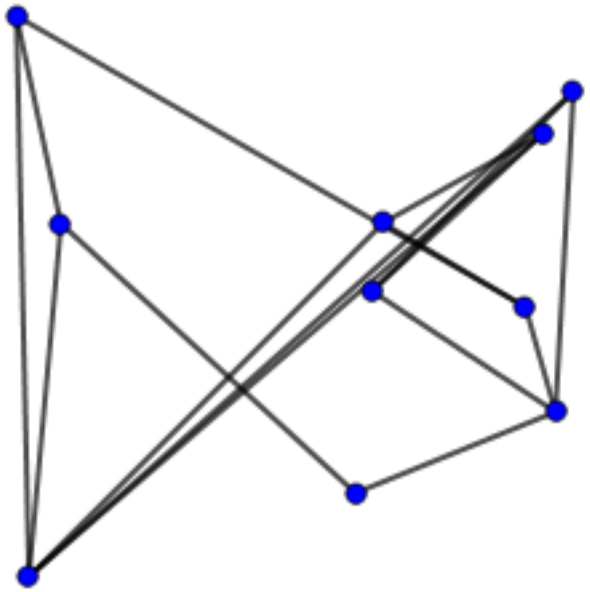
Example

- Draw n = 10 from prior ρ^0
- Let $p^0(x)$ be a 2-d unit Gaussian



Example

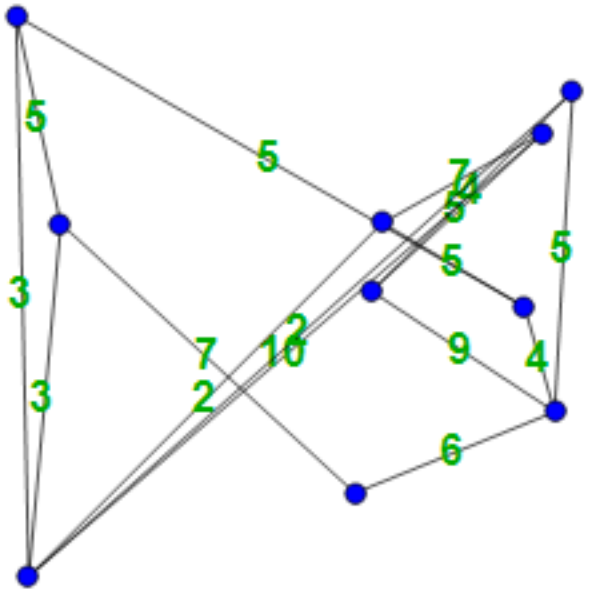
- Draw n = 10 from prior ρ^0
- Let $p^0(x)$ be a 2-d unit Gaussian
- Draw trading partner graph from $\mathcal{G}(10,0.4)$ (i.e., p = 0.4)



METRON
SCIENTIFIC SOLUTIONS

Example

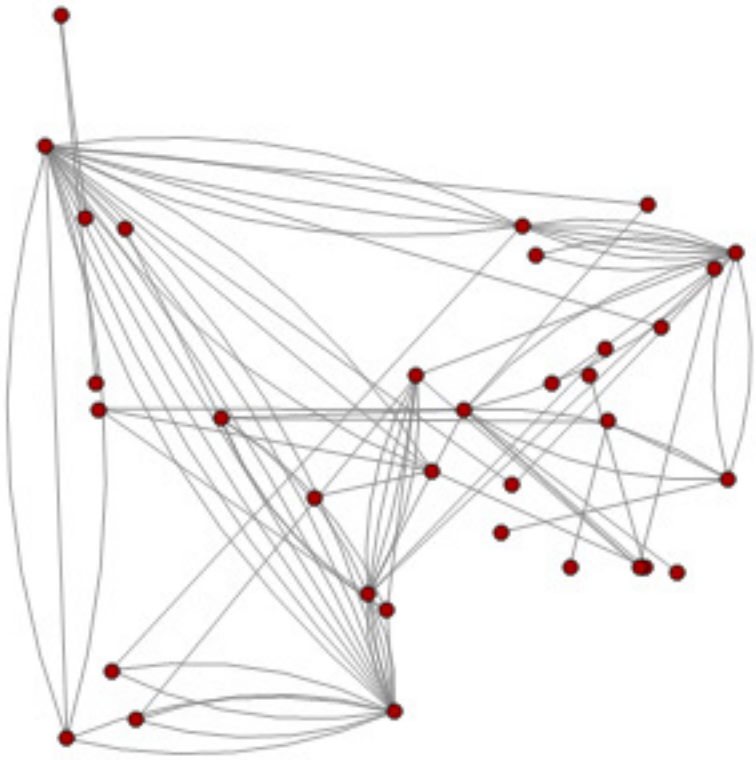
- Draw n = 10 from prior ρ^0
- Let $p^0(x)$ be a 2-d unit Gaussian
- Draw trading partner graph from $\mathcal{G}(10,0.4)$ (i.e., p = 0.4)
- Draw the number of transactions between trading partners from $\mathrm{Po}(\lambda t)$ with λt = 5



METRON
SCIENTIFIC SOLUTIONS

Example

- Partition *stubs* for node of degree d using $\mathrm{CRP}(d;\theta)$ with $\theta = 1$
  - Results are clusters of stubs for each node
- Draw alias z for each cluster of stubs
  - If entity's state is x
  - Draw alias z from $\mathcal{N}(x,\sigma I)$
  - With $\sigma = 0.4$
- What is the probability of an association (i.e., partition) given the pictured evidence?



METRON
SCIENTIFIC SOLUTIONS

Sketch of Derivation

- Ground truth: $X = (E^+, \mathbf{x}, n, \vec{p})$
  - E^+ = multiset of ground truth trading partners + trade multiplicities
  - $\mathbf{x}$ = array of states of all n entities
  - n = number of entities
  - $\vec{p}$ = vector of any parameters we wish to integrate out
- Prior probability:

$$P(X) = P(E^+ \mid n, \vec{p}) P(n \mid \vec{p}) \prod_{j=1}^{n} P(x_j)$$

- Data: $Z = (T^+, \mathbf{z})$
  - T^+ = multiset of observed transactions
  - $\mathbf{z}$ = array of all m aliases
- Assignment function a
  - $a(i) = j$ maps an alias index i to its entity's index j

METRON SCIENTIFIC SOLUTIONS

Sketch of Derivation

- After some work we find

$$P(T^+, \mathbf{z}, a \mid E^+, \mathbf{x}) = \frac{\theta^m}{m!} \prod_{e \in E} k_e ! \prod_{i=1}^{m} \Gamma(d_i) \prod_{j=1}^{n} \left(\frac{\Gamma(\theta)}{\Gamma(\theta + d_j)} \prod_{i \in a^{-1}(j)} P(z_i \mid x_j) \right)$$

  - where k_e is the number of transactions on ground truth edge e, and
  - d_i is the degree of the alias (z_i) indexed by i, whereas
  - d_j is the degree of the entity (x_j) indexed by j
- We integrate this against the priors on E^+, $\mathbf{x}$, etc.
- Once $\mathbf{x}$ is gone, what does assignment function a mean?
  - Group mappings a (to labeled entities) into partition A = $[a]$ of aliases
  - Introduces a combinatorial factor that counts mappings
- Integrate out as much as possible
  - Poisson rate λt
  - Erdős–Rényi p, and
  - Number of entities n

METRON SCIENTIFIC SOLUTIONS

Result: Graph Component

- We find

$$P(A \mid T^+, \mathbf{z}) \propto \left(\sum_{n=n_0}^{\infty} F(n) \right) \prod_{\alpha \in A} \frac{\Gamma(\theta) P(\mathbf{z}_\alpha)}{\Gamma(\theta + d_\alpha)}$$

  - where

$$F(n) = \frac{(n-1)!}{(n-n_0)!} \frac{1}{\mathrm{B}(\delta, n)} \int_0^\infty \int_0^1 p^{\delta-1} (1-p)^{n-1} (pe^{-\lambda})^{|E_0|} \left(1 - (1-e^{-\lambda})p\right)^{\binom{n}{2} - |E_0|} \lambda^{|T^+|-1} \, dp d\lambda$$

  - is well approximated by the lower bound

$$F^*(n) = \frac{\Gamma(|T^+|)}{|E_0|^{|T^+|}} \varphi(n) \quad \text{with} \quad \varphi(n) = \frac{(n-1)!}{(n-n_0)!} \frac{\mathrm{B}(\delta + |E_0|, n(n+1)/2 - |E_0|)}{\mathrm{B}(\delta, n)}$$

- Here n_0 and $|E_0|$ are the number of nodes and edges in the (graph-homomorphic) core induced by the association A

METRON SCIENTIFIC SOLUTIONS

Result: Spatial Component

- In

$$P(A \mid T^{+}, \mathbf{z}) \propto \left(\sum_{n=n_0}^{\infty} F(n) \right) \prod_{\alpha \in A} \frac{\Gamma(\theta) P(\mathbf{z}_\alpha)}{\Gamma(\theta + d_\alpha)}$$

  - the spatial component is

$$P(\mathbf{z}_\alpha) = (2\pi w_\alpha)^{r/2} \mathcal{N}(\mu_\alpha; 0, I) \prod_{i \in \alpha} \mathcal{N}(z_i; \mu_\alpha, \sigma^2 I)$$

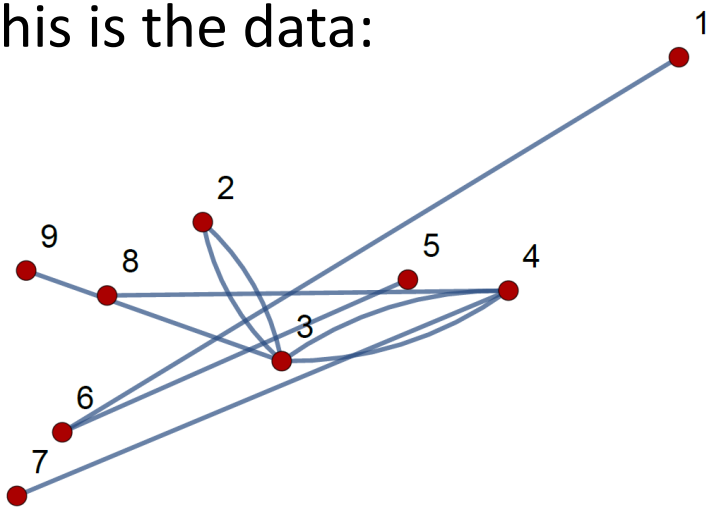
  - where

$$w_\alpha = \frac{\sigma^2}{|\alpha| + \sigma^2} \quad \text{and} \quad \mu_\alpha = \frac{|\alpha| \bar{z}_\alpha + \sigma^2}{|\alpha| + \sigma^2} \quad \text{with} \quad \bar{z}_\alpha = \frac{1}{|\alpha|} \sum_{i \in \alpha} z_i$$

METRON SCIENTIFIC SOLUTIONS

Numerical Example

- This is the data:

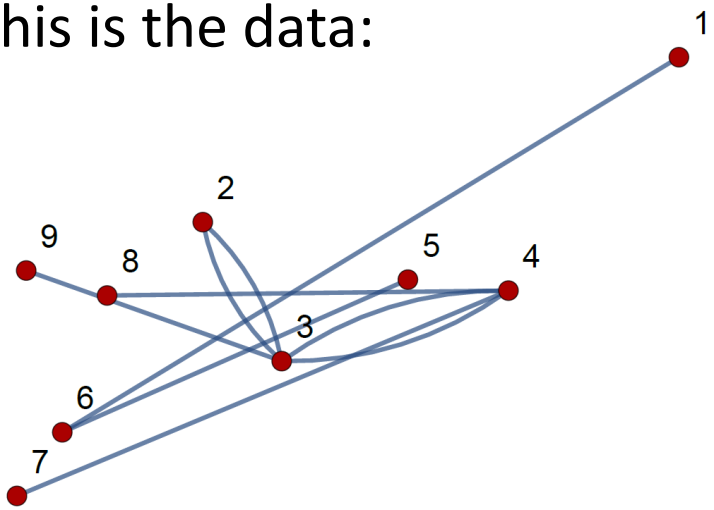


- Which aliases came from the same entity?
  - Hmm… 1 is its own thing (spatially)
  - 2 and 9 are both connected only to 3 (and are fairly close)
  - Any two connected aliases must be from different entities
  - 6 and 7 look like a group (spatially)
  - If 6 and 7 are the same, it's more evidence 4 and 5 are the same…

METRON SCIENTIFIC SOLUTIONS

Top 10 Associations

- This is the data:



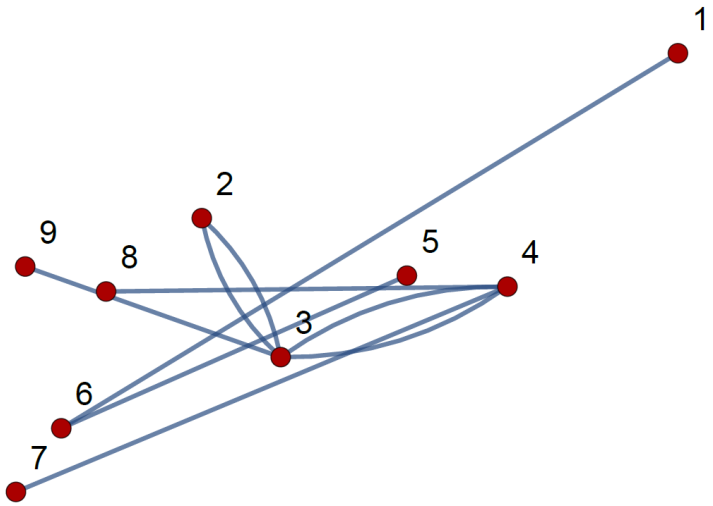
TEN MOST PROBABLE ASSOCIATIONS. ✓ = TRUTH.

	Pr(A)	A
→	0.250	{{1}, {2, 9}, {3}, {4, 5}, {6, 7, 8}}
	0.111	{{1}, {2, 9}, {3, 8}, {4, 5}, {6, 7}}
	0.072	{{1}, {2, 8, 9}, {3}, {4, 5}, {6, 7}}
	0.059	{{1}, {2, 9}, {3}, {4, 5}, {6, 7}, {8}}
	0.048	{{1, 4, 5}, {2, 9}, {3}, {6, 7, 8}}
✓	0.048	{{1}, {2}, {3}, {4, 5}, {6, 7, 8}, {9}}
	0.034	{{1}, {2}, {3}, {4, 5}, {6, 7}, {8, 9}}
	0.024	{{1, 4}, {2, 9}, {3}, {5}, {6, 7, 8}}
	0.023	{{1}, {2}, {3}, {4, 5}, {6, 7, 8, 9}}
	0.021	{{1, 4, 5}, {2, 9}, {3, 8}, {6, 7}}

- Which aliases came from the same entity?
  - Let's see… 1 is typically its own thing
  - 2 and 9 are probably from the same entity (but in truth: no! <gasp!>)
  - Any two connected aliases must be from different entities (yep)
  - 6 and 7 arise from the same entity in all 10 of the top associations
  - And 4 and 5 arise from the same entity in 9 out of 10

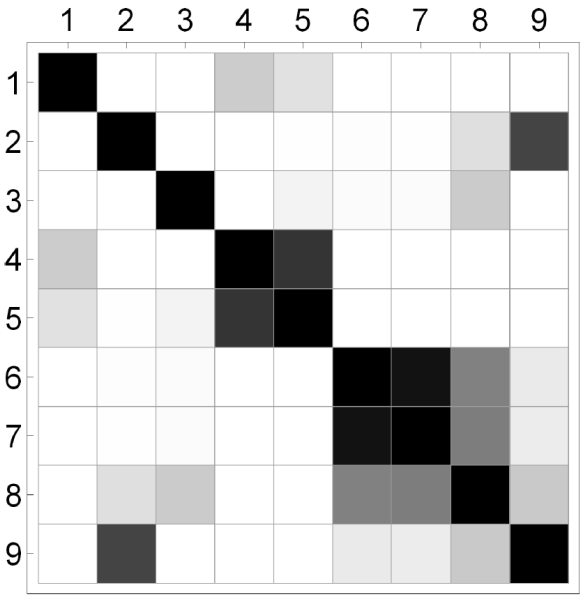
METRON SCIENTIFIC SOLUTIONS

Pairwise Co-membership Probabilities



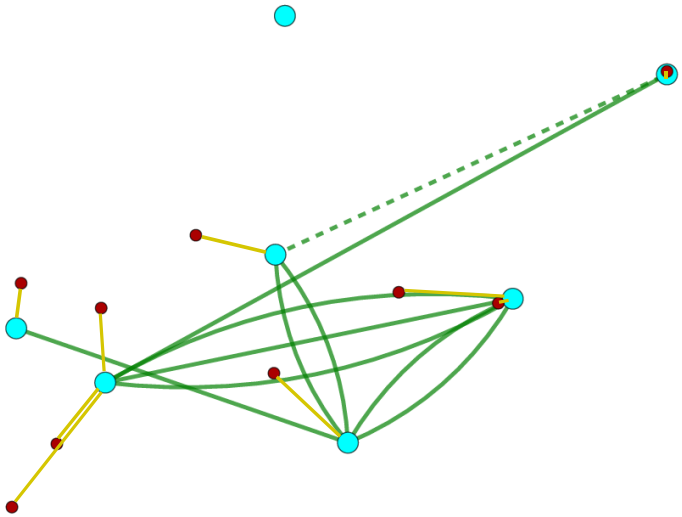
TEN MOST PROBABLE PAIRS. ✓ = TRUTH.

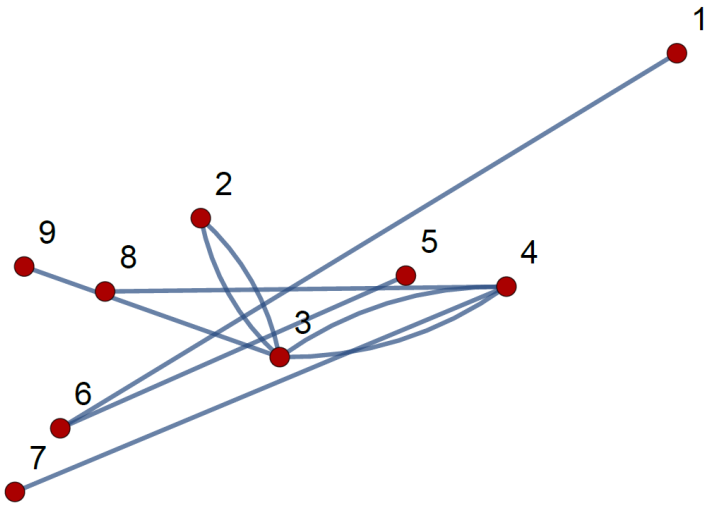
	$\Pr(\{i, i'\})$	$\{i, i'\}$
✓	0.927	{6, 7}
✓	0.795	{4, 5}
	0.733	{2, 9}
✓	0.509	{7, 8}
✓	0.494	{6, 8}
	0.211	{8, 9}
	0.203	{3, 8}
	0.202	{1, 4}
	0.125	{2, 8}
	0.117	{1, 5}



METRON
SCIENTIFIC SOLUTIONS

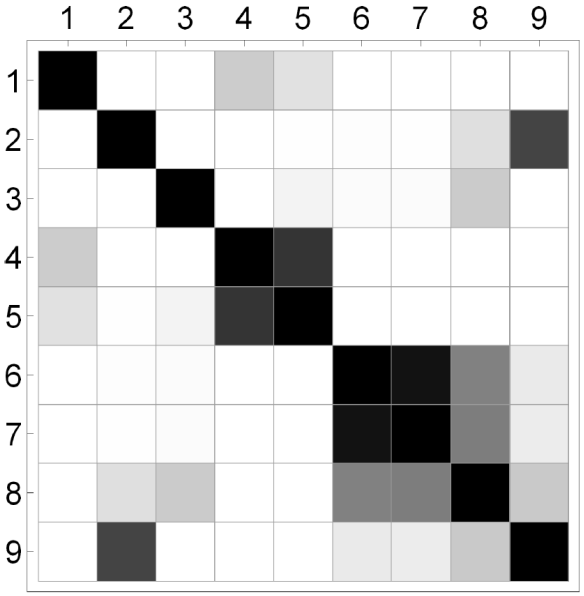
Pairwise Co-membership Probabilities





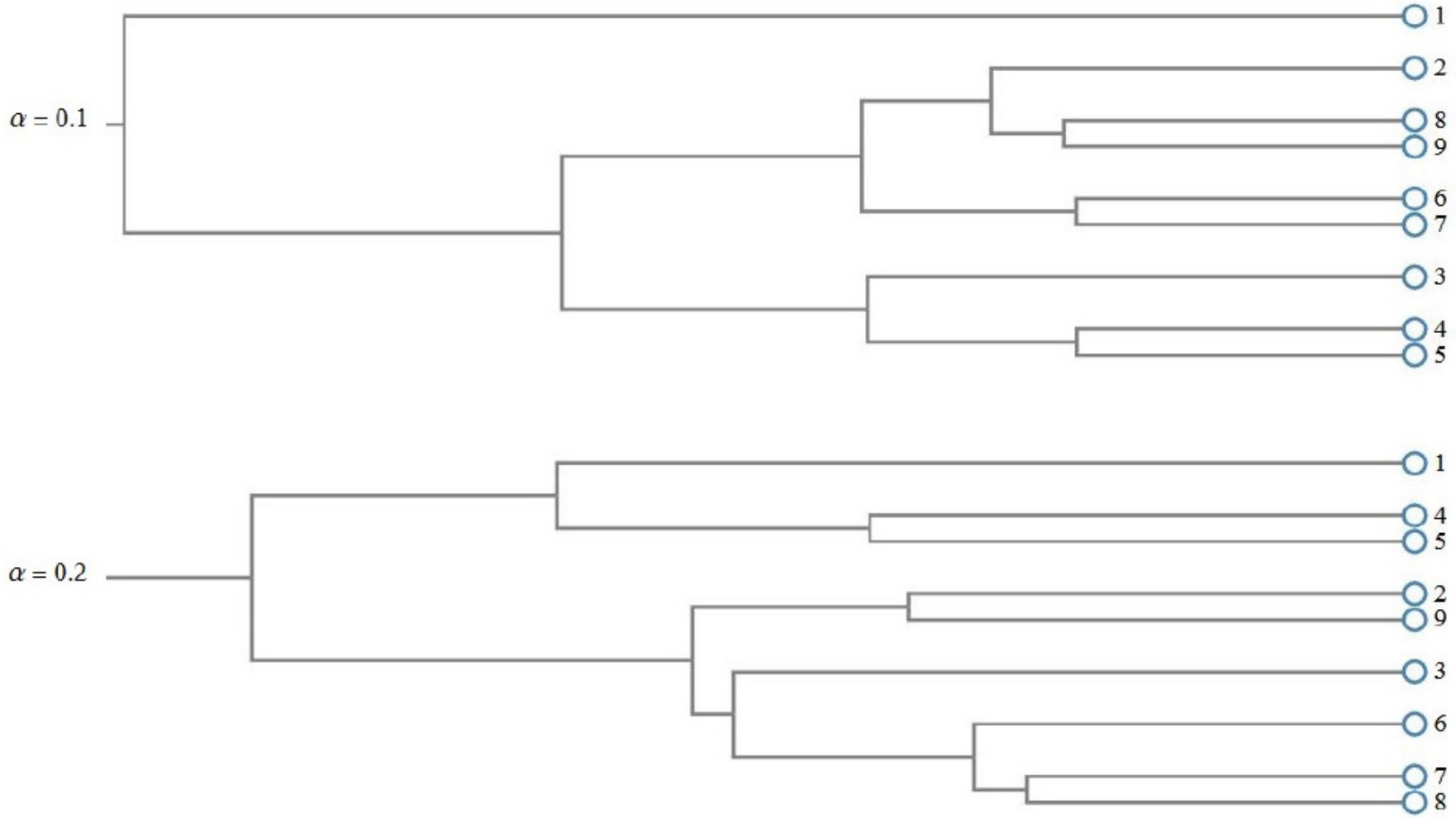
TEN MOST PROBABLE PAIRS. ✓ = TRUTH.

	$\Pr(\{i, i'\})$	$\{i, i'\}$
✓	0.927	{6, 7}
✓	0.795	{4, 5}
	0.733	{2, 9}
✓	0.509	{7, 8}
✓	0.494	{6, 8}
	0.211	{8, 9}
	0.203	{3, 8}
	0.202	{1, 4}
	0.125	{2, 8}
	0.117	{1, 5}



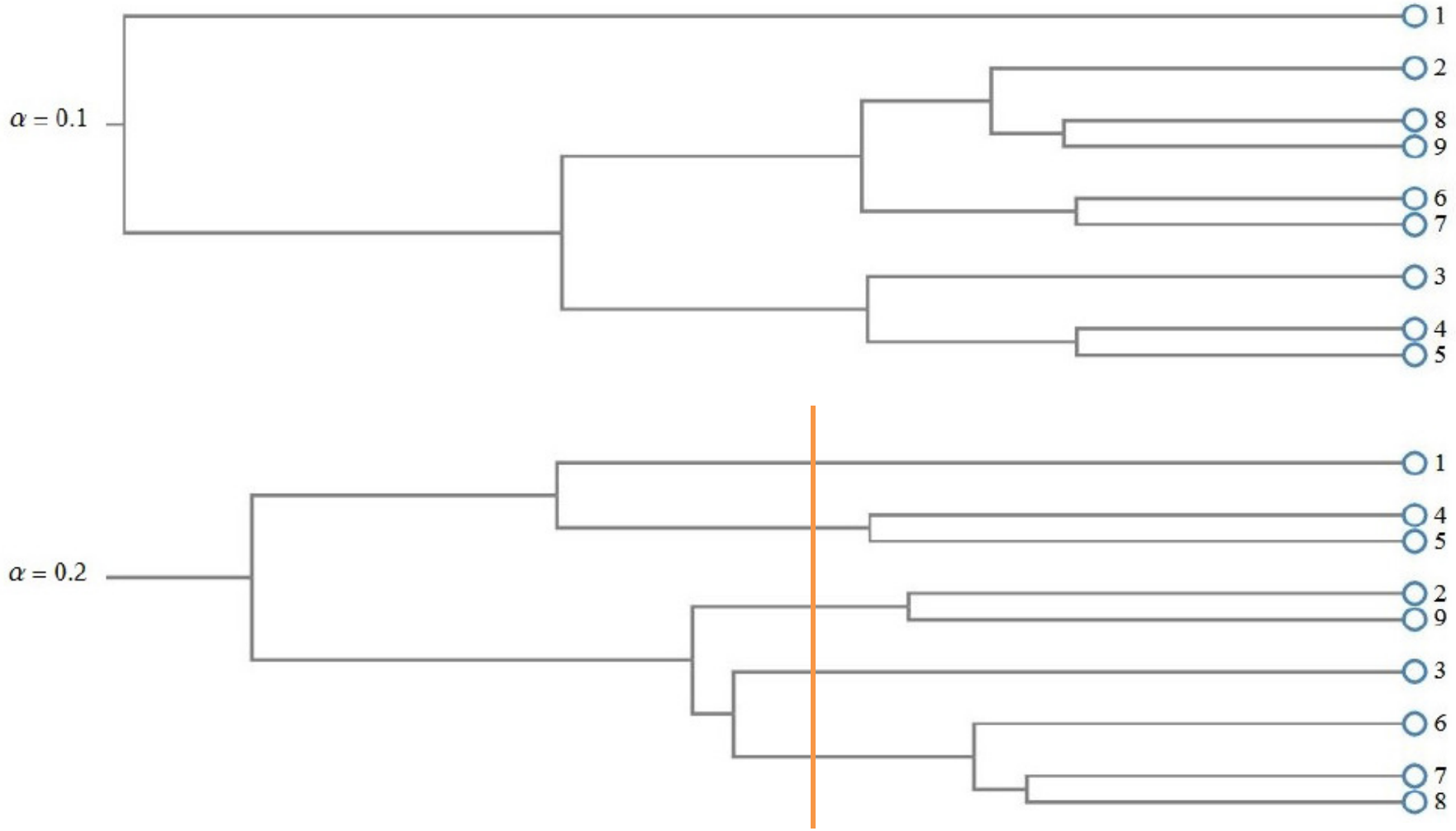
METRON
SCIENTIFIC SOLUTIONS

Comparison to Bhattacharya–Getoor Algorithm



METRON
SCIENTIFIC SOLUTIONS

Comparison to Bhattacharya–Getoor Algorithm



METRON
SCIENTIFIC SOLUTIONS

Summary

- Entity Resolution
  - Good algorithms exist, but
  - This is the first generative model we know of, and
  - It's useful for Small Data (no association hypothesis a clear winner)
- Mathematical Idealizations
  - Why study the Erdős–Rényi $\mathcal{G}(n,p)$ model *that much*?
    - It's not realistic…
  - It's compelling for its own sake
  - Overriding question: which practical problems have a mathematical structure that's compelling for its own sake?
    - The Eternal Search for Truth and Beauty, or
    - Tricking mathematicians into doing something useful
- Thank you to my co-authors: Darren Lo and Thomas Seaquist
  - J. P. Ferry, D. Lo, and T. Seaquist, *A Bayesian Idealization of Entity Resolution*, Proceedings of the 18th International Conference on Information Fusion, Washington DC, July 6-9, 2015.

METRON SCIENTIFIC SOLUTIONS