Structure of the talk

- Finding links and initiators: a graph-reconstruction problem

  H. Mannila & E. Terzi, SDM 2009

	Finding experts	Finding links and initiators
Network structure	**Known**	**Unknown**
Nodes' features	**Known**	**Known**

Why care?

- Which was the site in which a species first appeared?
- How do species migrate from site to site?

Why care?

David				
Cynthia				
Bob				
Alice				

- Who introduced the Nokia phone?
- Who introduced the iPhone?
- How are Alice, Bob, Cynthia and David influencing each other's purchases?

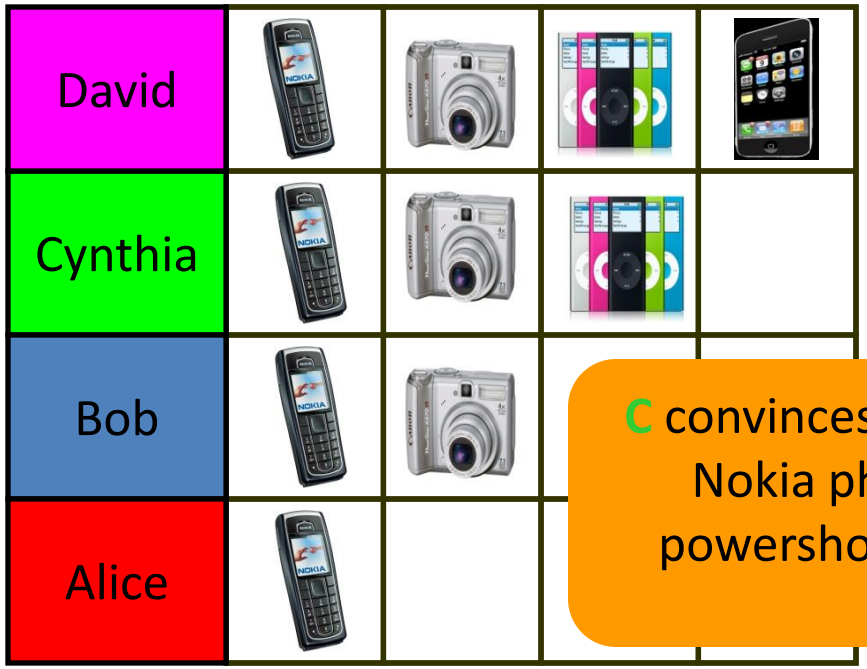
Framework

- Data (D): 0/1 (presence/absence) matrices of signals (columns) that appear to entities (rows)
- Given D [$n_x m$] and a propagation model of signals from one entity to another find:
  - **Connections/links** between **entities (G [$n_x n$])**
  - **Initiator entities** for the different signals (**I [$n_x m$]**)

Applications

- Ecology
  - **Input:** Presence/absence matrices for species and sites
  - **Output:** Migration patterns of species across sites + sites where species first appeared
- Social networks/Customer transactions
  - **Input:** Transactions of customers “who bought what”
  - **Output:** Social network of customers + inference of customers that created trends (initiators)

Links and initiators



C convinces D to buy Nokia phone + powershot + ipod

B convinces C to buy Nokia phone + powershot

David

D initiates iPhone

Cynthia

C initiates iPod

Bob

B initiates Powershot

Alice

A initiates Nokia phone

Links and initiators

D=

David				
Cynthia				
Bob				
Alice				

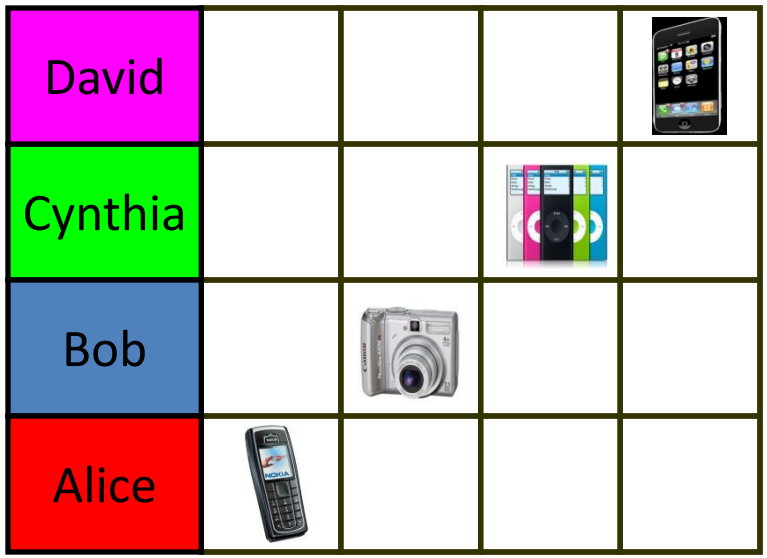
I =

David				
Cynthia				
Bob				
Alice				

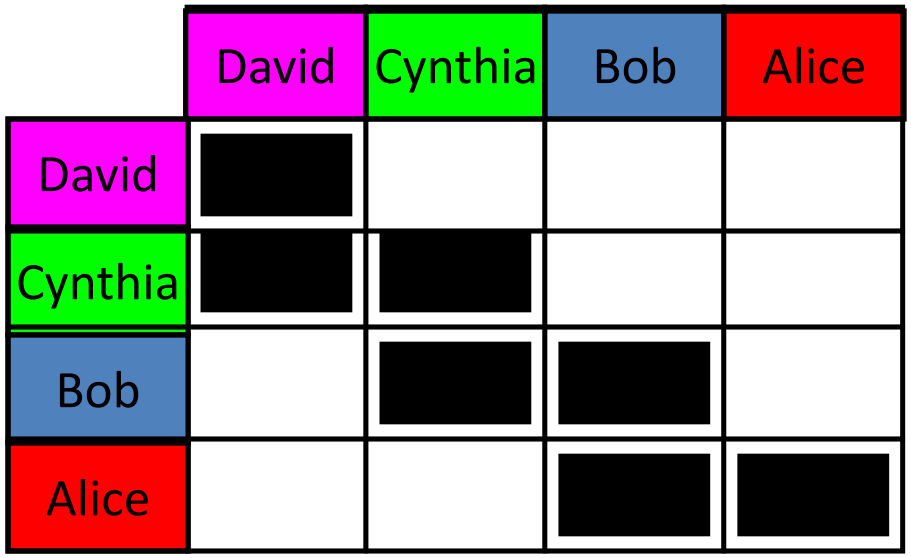
G =

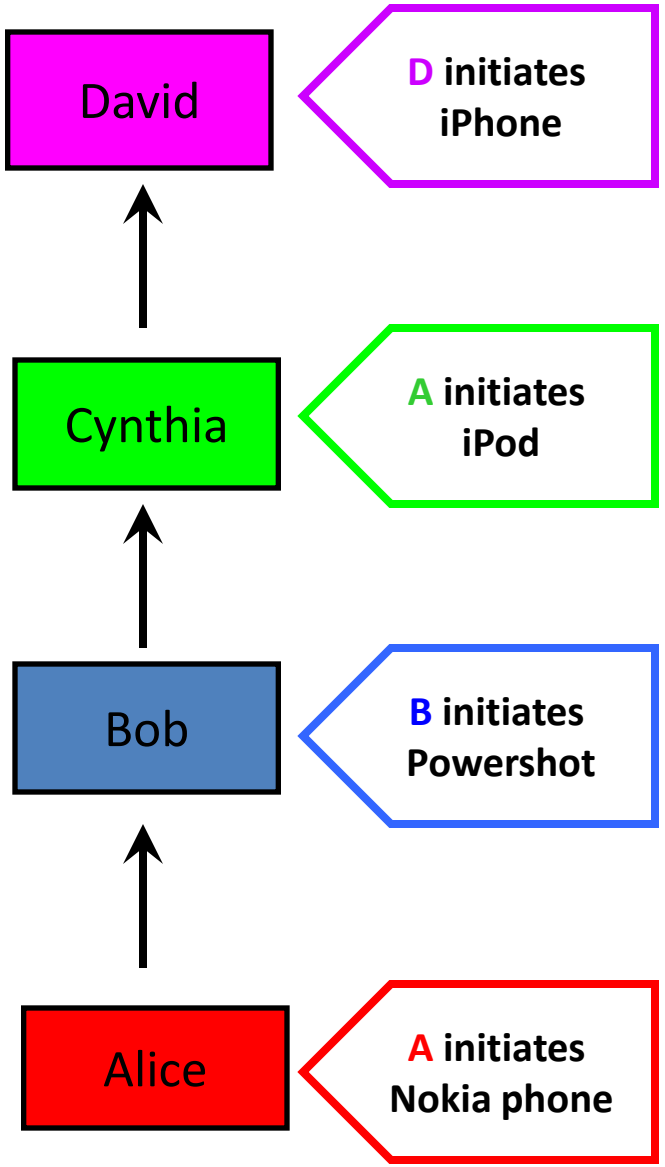
	David	Cynthia	Bob	Alice
David				
Cynthia				
Bob				
Alice				

I =


David
Cynthia
Bob
Alice

G =

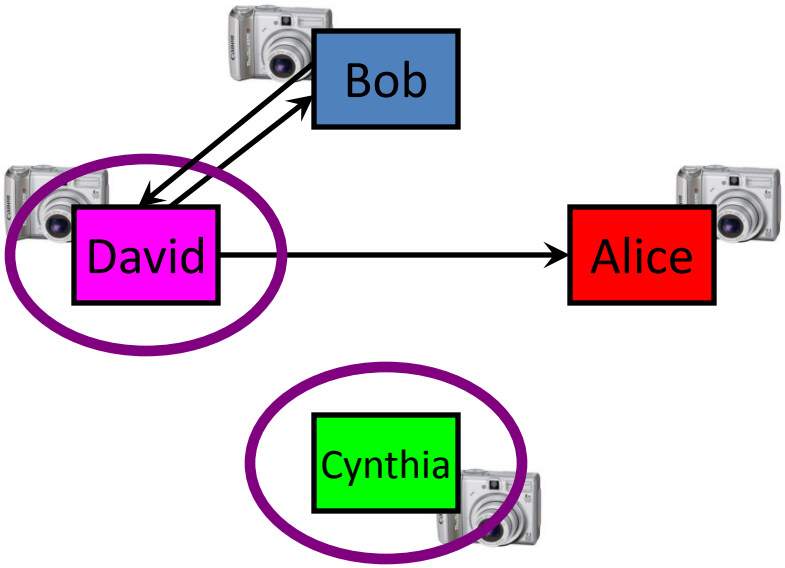

David
Cynthia
Bob
Alice
David
Cynthia
Bob
Alice

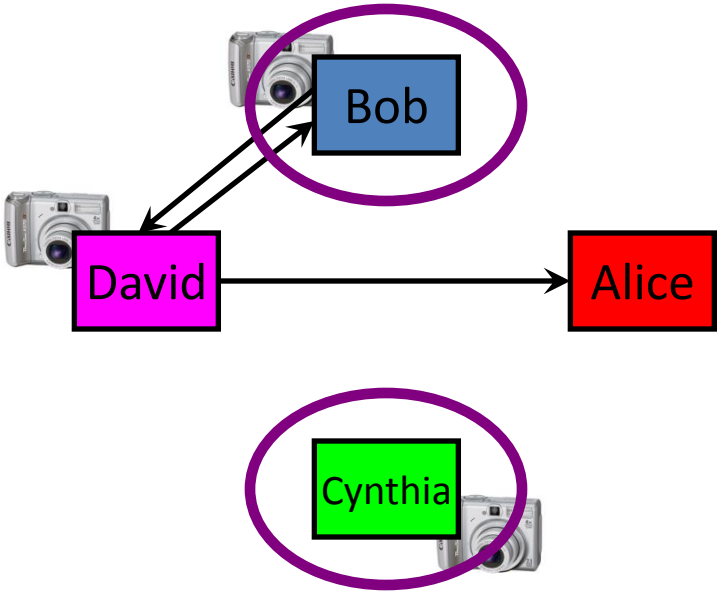

David
D initiates iPhone
Cynthia
A initiates iPod
Bob
B initiates Powershot
Alice
A initiates Nokia phone

How to find links and **initiators**?

Influence depends on the length of the path connecting two nodes in the graph

- Assume we know the graph (links) and we only want to find the initiators





Finding k initiators is NP-hard

Finding k initiators is NP-hard and NP-hard to approximate!

How to find links **and** initiators?

- Our problem is even harder!!
  - We don't know the initiators
  - Neither do we know the graph
- Candidate problem definition: Given observation matrix D find graph-initiator pair G, I such that

$$\{G,I\} = \text{argmax}_{\{G',I'\}} \Pr(G',I' \mid D)$$

Posterior probability of graph-initiator pair

Disadvantages
The best solution is:
→G has no edges
→ I=D

Posterior probability of a graph-initiator pair

$$\Pr(G, I \mid D) \propto \Pr(D \mid G, I)\Pr(G)\Pr(I)$$

D: observation matrix [nxm]
G: social graph (directed) [nxn]
I: initiator matrix [nxm]

Penalize for complex G: Pr(G) = exp{-c1|E|}
Penalize for complex I : Pr(I) = exp{-c2|I|}

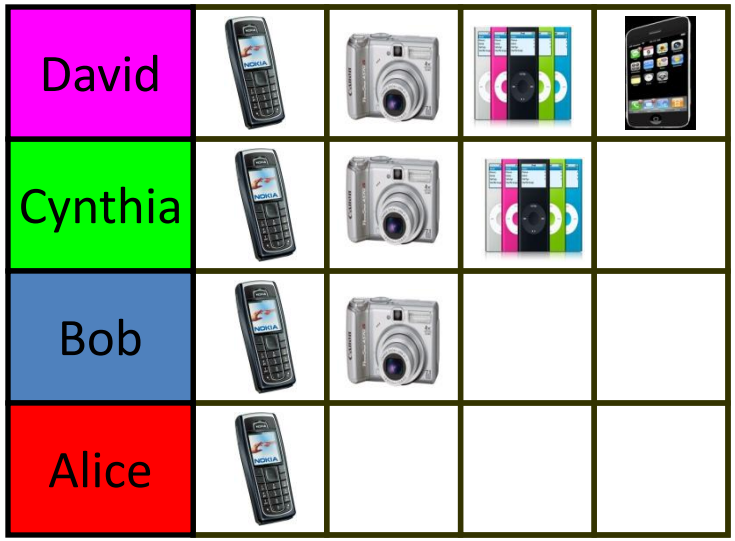
- **$\Pr(D|G,I) = \prod_{i=1\ldots n} \prod_{u=1\ldots m} \Pr(D_{(i,u)}|G,I)$**
- **D(i,u) = 0**
  - $\Pr(D_{(i,u)}=0|G,I) = (1-I(i,u)) \quad x \quad \prod_{j\neq i}(1-I(j,u))\, b(j,i,G)$
- **D(i,u) = 1**
  - $\Pr(D_{(i,u)}=1|G,I) = 1 - \Pr(D_{(i,u)}=0|G,I)$

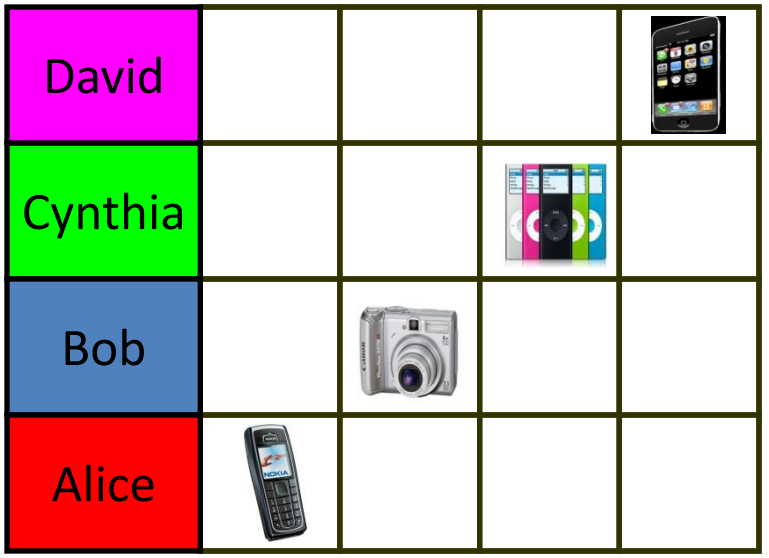
Problem definition v.2

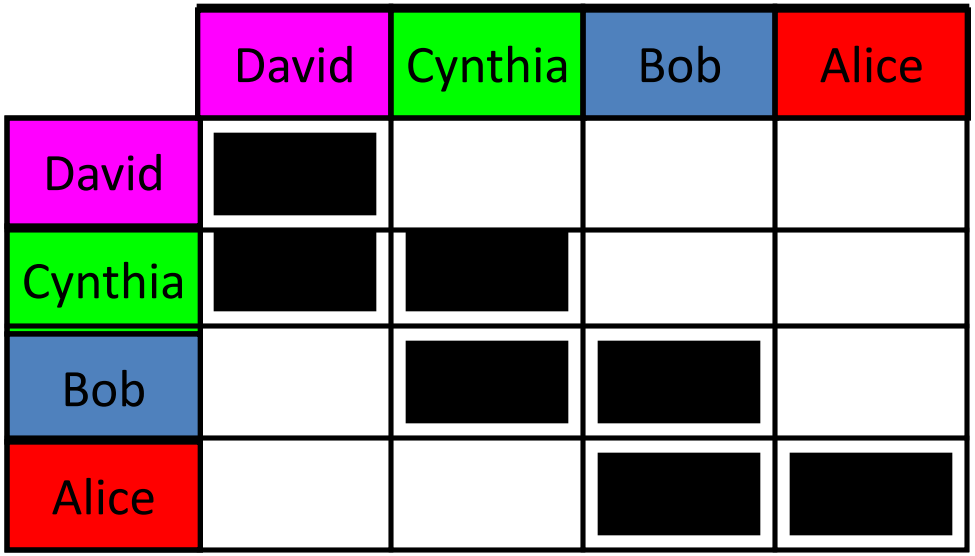
- **Given observation matrix D find graph initiator pair G, I such that**

$$\{G,I\} = \text{argmax}\{G',I'\} \Pr(D|G',I')\Pr(G')\Pr(I')$$

No unique solution







No unique solution

David				
Cynthia				
Bob				
Alice				

David				
Cynthia				
Bob				
Alice				

	David	Cynthia	Bob	Alice
David				
Cynthia				
Bob				
Alice				

Sampling

- There is no unique good solution to the maximum likelihood problem
  - The number of bits of the input are less than the number of bits in the output
- Sample the space of solutions instead of finding a unique solution

The MCMC algorithm

- Sample the space of graph initiator pairs {G,N}

- Start with a random pair {G,N}
- For i = 1 to N do
  - {G',I'} = LocalMove ({G,I})
  - {G,I} = {G',I'} with Prob = min{1,Pr(G',I'|D)/Pr(G,I|D)}

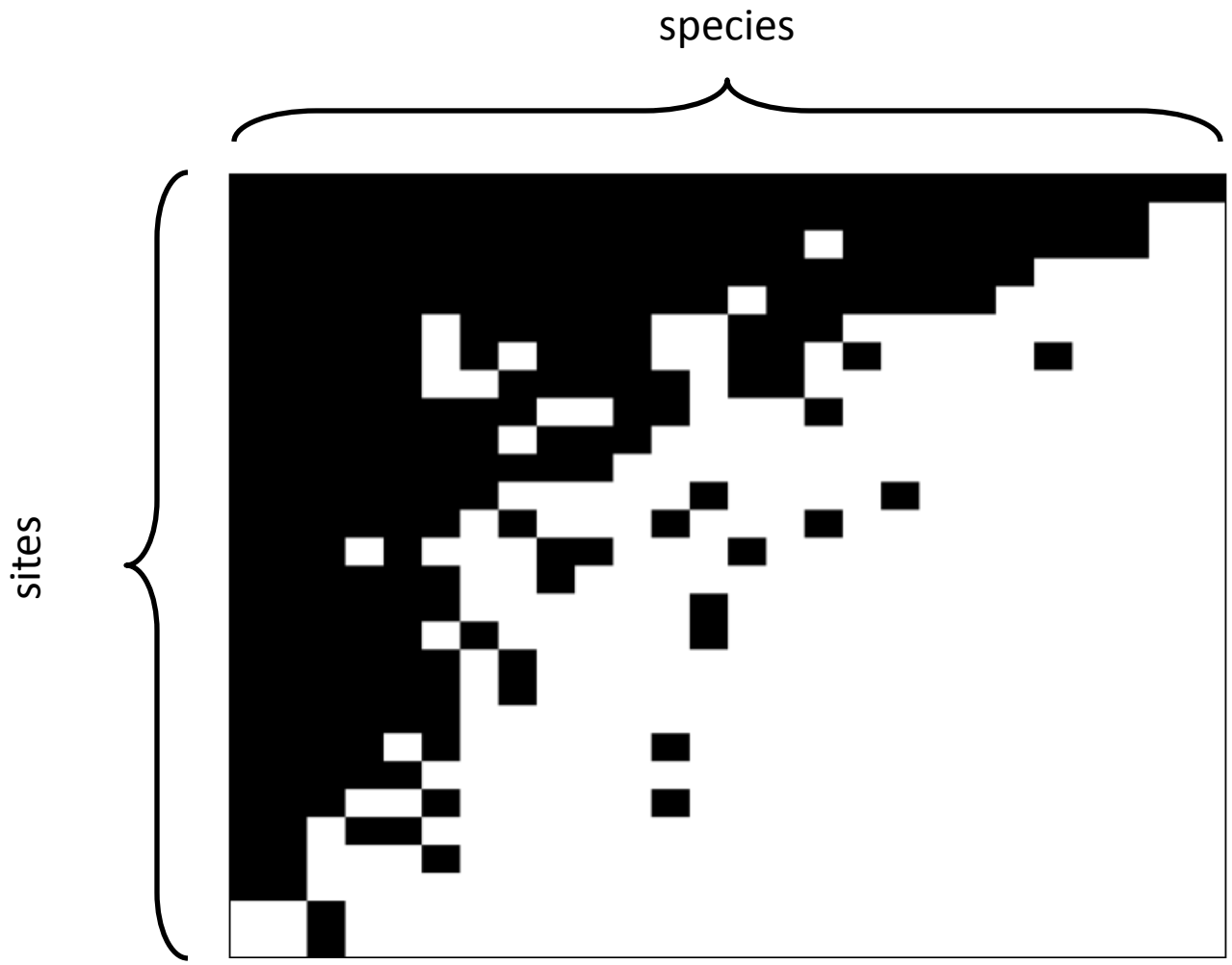
Flip entries in G and N

- Count how many times an edge exists and a client initiates a product – report these counts

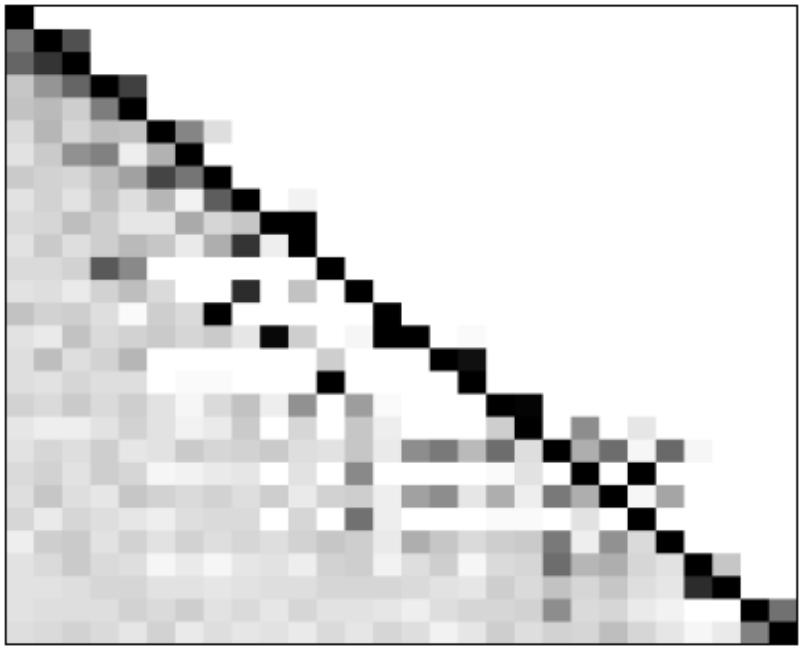
Incorporating temporal information

- One can take into account an ordered sequence of observation matrices $D_1, D_2, \ldots, D_k$
- Framework extents to this setting easily
  - Just a more difficult way to compute $\mathbf{Pr(G,I \mid D_1,\ldots,D_k)}$

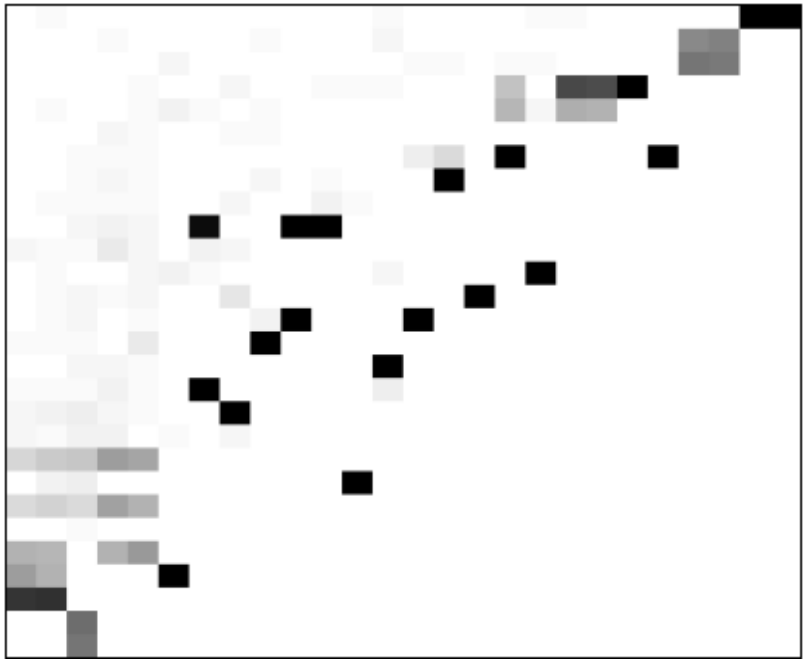
Ecological dataset: Rocky Mountain



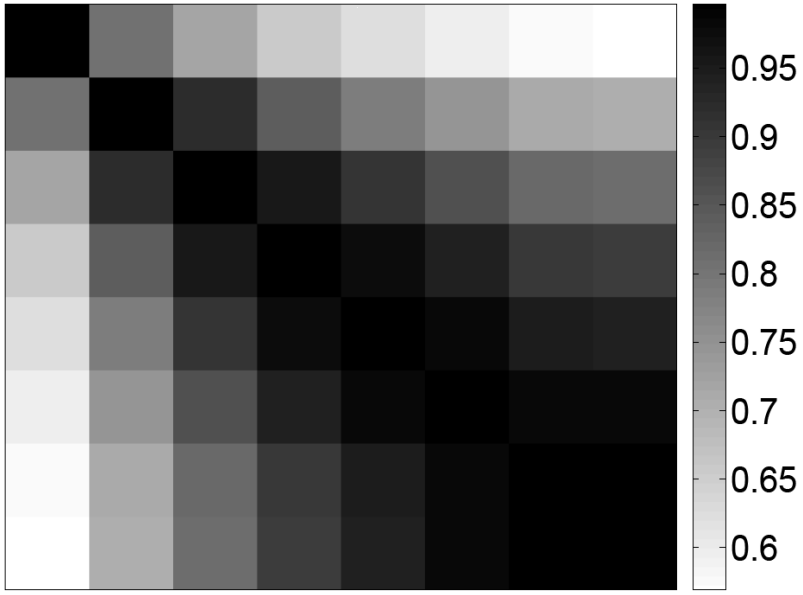
Ecological dataset: links between sites inferred by MCMC



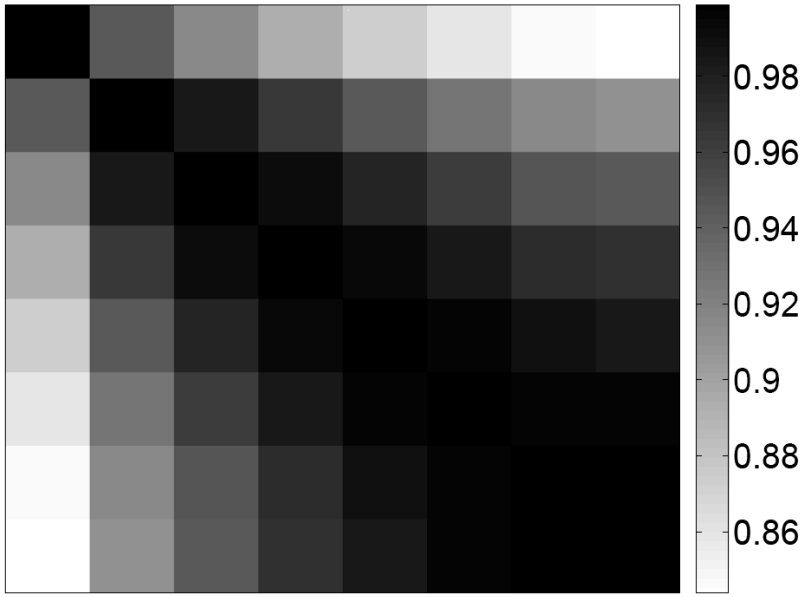
Ecological dataset: initiators inferred by MCMC



Ecological dataset: convergence



links



initiators