Flows on Random networks: Probabilistic and Statistical Problems

Recent Advances in probability
ISI

Shankar Bhamidi

Statistics Department
U.C. Berkeley

December 13th, 2007

Motivation: Present day context

First wave: Statistical Physicists

- Availability of data from real world networks like the Internet, various biological networks, spatial networks e.g. road and rail networks.

First wave: Statistical Physicists

- Availability of data from real world networks like the Internet, various biological networks, spatial networks e.g. road and rail networks.
- Caused an explosion in the number of network models largely by Statistical Physicists to explain various characteristics of "'‘real world networks"’.

Motivation: Present day context

First wave: Statistical Physicists

- Availability of data from real world networks like the Internet, various biological networks, spatial networks e.g. road and rail networks.
- Caused an explosion in the number of network models largely by Statistical Physicists to explain various characteristics of "'‘real world networks"'’.
- were instrumental (using e.g. branching processes) in making quantitative predictions about their models

Motivation: Present day context

First wave: Statistical Physicists

- Availability of data from real world networks like the Internet, various biological networks, spatial networks e.g. road and rail networks.
- Caused an explosion in the number of network models largely by Statistical Physicists to explain various characteristics of "'‘real world networks"'.
- were instrumental (using e.g. branching processes) in making quantitative predictions about their models
- also gave rise to 10 second sound bite science: "'‘Internet is Robust yet fragile. 95 % of the links can be removed and the graph will stay connected. However targeted removal of 2.3 % of the hubs would disconnect the internet."'

Probabilistic motivations contd.

Second wave: Theoreticians

- prove above conjectures regarding these models.
- via impetus from computer science study various random dynamics on these models
- random walks mixing times, epidemic models(spread of virus on the internet), efficiency of local search algorithms.

Aim of this talk

Explore one aspect of this vast field, namely flows on networks.

Plan of the Talk

1. Motivation: Why care ?

Plan of the Talk

1. Motivation: Why care ?
2. Betweeness Centrality and multicommodity flow

Plan of the Talk

1. Motivation: Why care ?
2. Betweeness Centrality and multicommodity flow
3. Math model on the Complete Graph.

Plan of the Talk

1. Motivation: Why care ?
2. Betweeness Centrality and multicommodity flow
3. Math model on the Complete Graph.
4. Network Tomography - Network inference via indirect measurements.

Plan of the Talk

1. Motivation: Why care ?
2. Betweeness Centrality and multicommodity flow
3. Math model on the Complete Graph.
4. Network Tomography - Network inference via indirect measurements.
5. Math Ideas for Betweeness centrality on complete graph.

Plan of the Talk

1. Motivation: Why care ?
2. Betweeness Centrality and multicommodity flow
3. Math model on the Complete Graph.
4. Network Tomography - Network inference via indirect measurements.
5. Math Ideas for Betweeness centrality on complete graph.
6. Conclusion

Main goal of our study

Probabilistic side

Looking for tractable math models and "flow" problems where we have a large number of interacting agents competing for the same resources. Interested in finding "curves " which would tell us things about the "congestion" in the network at least for "large" networks

Statistics

Want to understand the structure of real-world networks especially the Internet. Often not possible to do so directly. Can we, using a small number of "indirect " samples reconstruct the network?

Betweeness Centrality

Problem setup

- Connected network on n nodes. Assume shortest path between every pair of nodes unique.
- every node sends flow of volume a_n to every other node using the *least cost path*
- For ordered pair of nodes (s, t) let $\pi(s, t)$ denote the least cost path.

Betweeness Centrality

Problem setup

- Connected network on n nodes. Assume shortest path between every pair of nodes unique.
- every node sends flow of volume a_n to every other node using the *least cost path*
- For ordered pair of nodes (s,t) let $\pi(s,t)$ denote the least cost path.
- For any directed edge e the amount of flow passing through the edge is

$$F_n(e) = a_n \#\{\text{(s,t) pairs which use edge e }\}$$

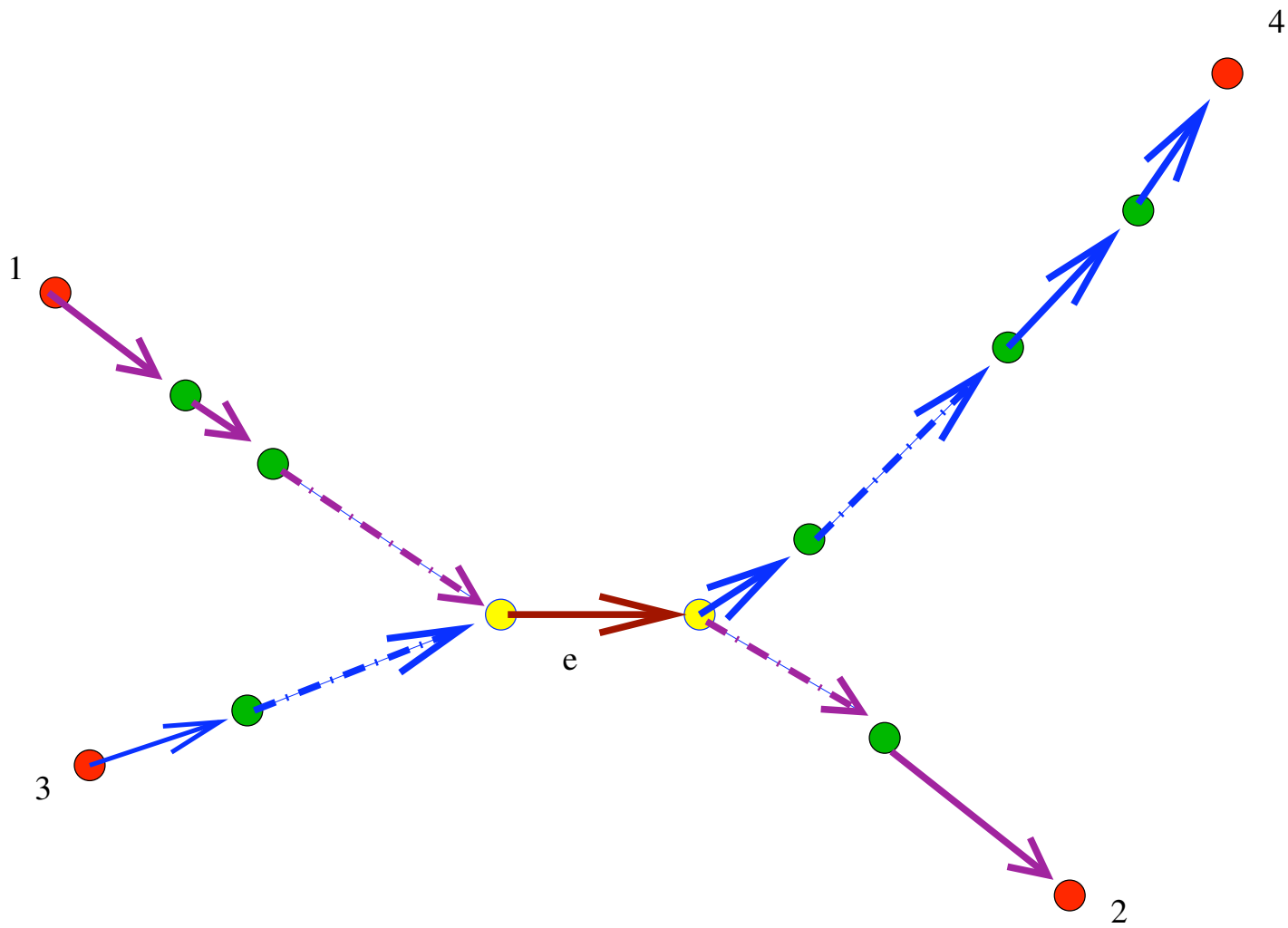


Figure 1: **Composite flow through edges**

Betweeness Centrality

Problem setup

- Connected network on n nodes. Assume shortest path between every pair of nodes unique.
- every node sends flow of volume a_n to every other node using the *least cost path*
- For ordered pair of nodes (s,t) let $\pi(s,t)$ denote the least cost path.
- For any directed edge e the amount of flow passing through the edge is

$$F_n(e) = a_n \#\{\text{(s,t) pairs which use edge e }\}$$

Math question:

Can we get asymptotics for the random empirical distribution of this edge flow namely

$$\#\{e : F_n(e) > z\} \tag{1}$$

1. Edge flows on the complete graph

Problem setup

- Take complete graph. Attach exponential edge costs(lengths) IID with mean 1.
- every node sends a flow of volume $1/n$ to every other node using the *least cost path*
- For ordered pair of nodes (s, t) let $\pi(s, t)$ denote the least cost path

1. Edge flows on the complete graph

Problem setup

- Take complete graph. Attach exponential edge costs(lengths) IID with mean 1.
- every node sends a flow of volume $1/n$ to every other node using the *least cost path*
- For ordered pair of nodes (s,t) let $\pi(s,t)$ denote the least cost path
- For any directed edge e the amount of flow passing through the edge is

$$F_n(e) = \frac{1}{n} \sum_{s,t} 1_{\{e \in \pi(s,t)\}}$$

1. Edge flows on the complete graph

Problem setup

- Take complete graph. Attach exponential edge costs(lengths) IID with mean 1.
- every node sends a flow of volume $1/n$ to every other node using the *least cost path*
- For ordered pair of nodes (s,t) let $\pi(s,t)$ denote the least cost path
- For any directed edge e the amount of flow passing through the edge is

$$F_n(e) = \frac{1}{n} \sum_{s,t} 1_{\{e \in \pi(s,t)\}}$$

Math question:

Can we get asymptotics for the random empirical distribution of this edge flow namely

$$\frac{1}{n} \#\{e : F_n(e) > z \log n\} \tag{2}$$

Complete graph

Changes in Geometry of Complete Graph

	without rand. costs	with rand. costs
# of edges on $\pi(s,t)$	1	$\log n$
typical flow	1/n	short edges carry $O(\log n)$

Complete graph

Changes in Geometry of Complete Graph

	without rand. costs	with rand. costs
# of edges on $\pi(s,t)$	1	$\log n$
typical flow	1/n	short edges carry $O(\log n)$

Complete Graph : n^2 edges.

$$\frac{1}{n}\#\{e : F_n(e) > z \log n\} \tag{3}$$

In the analysis typical "‘low cost"’ edges carry $O(\log n)$ flow and there are typically O(n) "‘low cost"’ edges.

Edge flows: Result (Joint work with David Aldous)

Result

As $n \to \infty$ for fixed $z > 0$,

$$\frac{1}{n}\#\{e : F_n(e) > z \log n\} \to_{L^1} G(z) \tag{4}$$

$G(z) = \int_0^\infty P(W_1 W_2 e^{-u} > z)\, du$
where W_1 and W_2 are independent Exponential(1).

Edge flows: Result (Joint work with David Aldous)

Result

As $n \to \infty$ for fixed $z > 0$,

$$\frac{1}{n}\#\{e : F_n(e) > z \log n\} \to_{L^1} G(z) \tag{4}$$

$G(z) = \int_0^\infty P(W_1 W_2 e^{-u} > z)\, du$
where W_1 and W_2 are independent Exponential(1).

Form of limiting G:

G(z) turns out to be asymptotic to

$$\sqrt{\pi} \exp(-(2\sqrt{z} + \frac{1}{4}\log z))$$

$$\#\{e : F_n(e) > z \log n\} \approx n\sqrt{\pi} \exp(-(2\sqrt{z} + \frac{1}{4}\log z))$$

Universality

Random demands

- above result independent of the demands between various nodes
- instead of $1/n$, take say D_{ij}/n, where D_{ij} independent with $\mathbb{E}(D_{ij}) = 1$, $\mathbb{E}(D_{ij}^2) < C$.
- Route everything via least cost paths

Universality

Random demands

- above result independent of the demands between various nodes
- instead of $1/n$, take say D_{ij}/n, where D_{ij} independent with $\mathbb{E}(D_{ij}) = 1$, $\mathbb{E}(D_{ij}^2) < C$.
- Route everything via least cost paths
- Same result (weak law of large numbers type result) namely

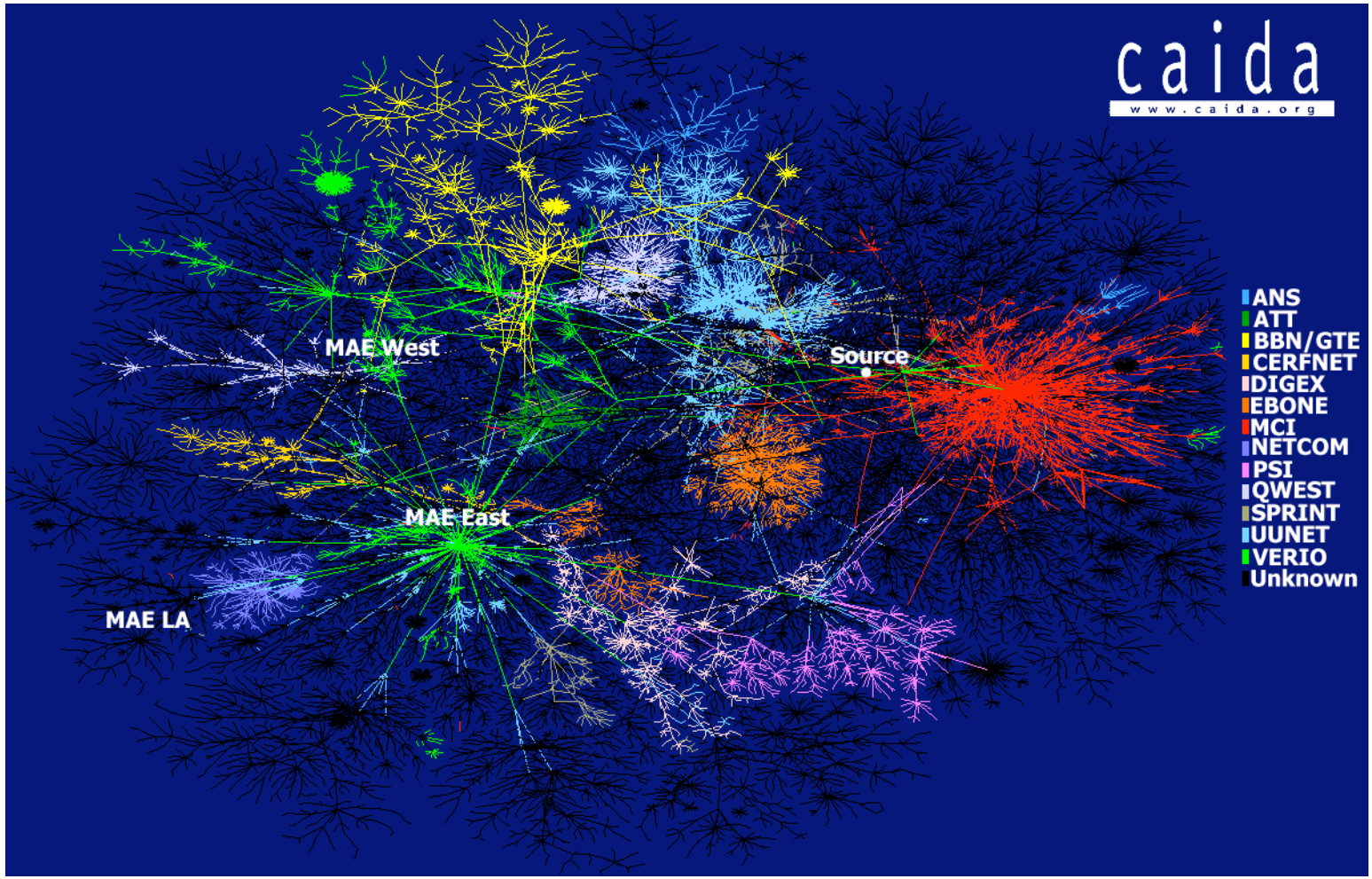
$$\frac{1}{n}\#\{e : F_n(e) > z \log n\} \longrightarrow_{L^1} G(z)$$

shall describe some of the details of the proof

Statistics: Network Tomography

(Joint Work with Ram Rajagopal and Sebastien Roch)

Why Network Inference necessary


caida
www.caida.org
MAE West
Source
MAE East
MAE LA
ANS
ATT
BBN/GTE
CERFNET
DIGEX
EBONE
MCI
NETCOM
PSI
QWEST
SPRINT
UUNET
VERIO
Unknown

Statistics: Network Tomography

(Joint Work with Ram Rajagopal and Sebastien Roch)

Why Network Inference necessary

Tomographic Image reconstruction

Tomography refers to the cross-sectional imaging of an object from either transmission or reflection data collected by illuminating the object from many different directions.

Statistics: Network Tomography

(Joint Work with Ram Rajagopal and Sebastien Roch)

Why Network Inference necessary

Tomographic Image reconstruction

Tomography refers to the cross-sectional imaging of an object from either transmission or reflection data collected by illuminating the object from many different directions.

Internal Approach

Measurements taken directly at the links or routers in the network

- extra computational burden at the routers
- congestion when transmitting collected data
- network operator may not allow access to these measurements

Network Tomography

External Approach

Uses "'‘end-end"’ measurements

Network Tomography

External Approach

Uses "'‘end-end"' measurements

Multicast Inference

- Single source allowed to send single data packet to some receivers via its routing tree.
- The packet duplicated at each branch point in the tree and sent further down the tree.

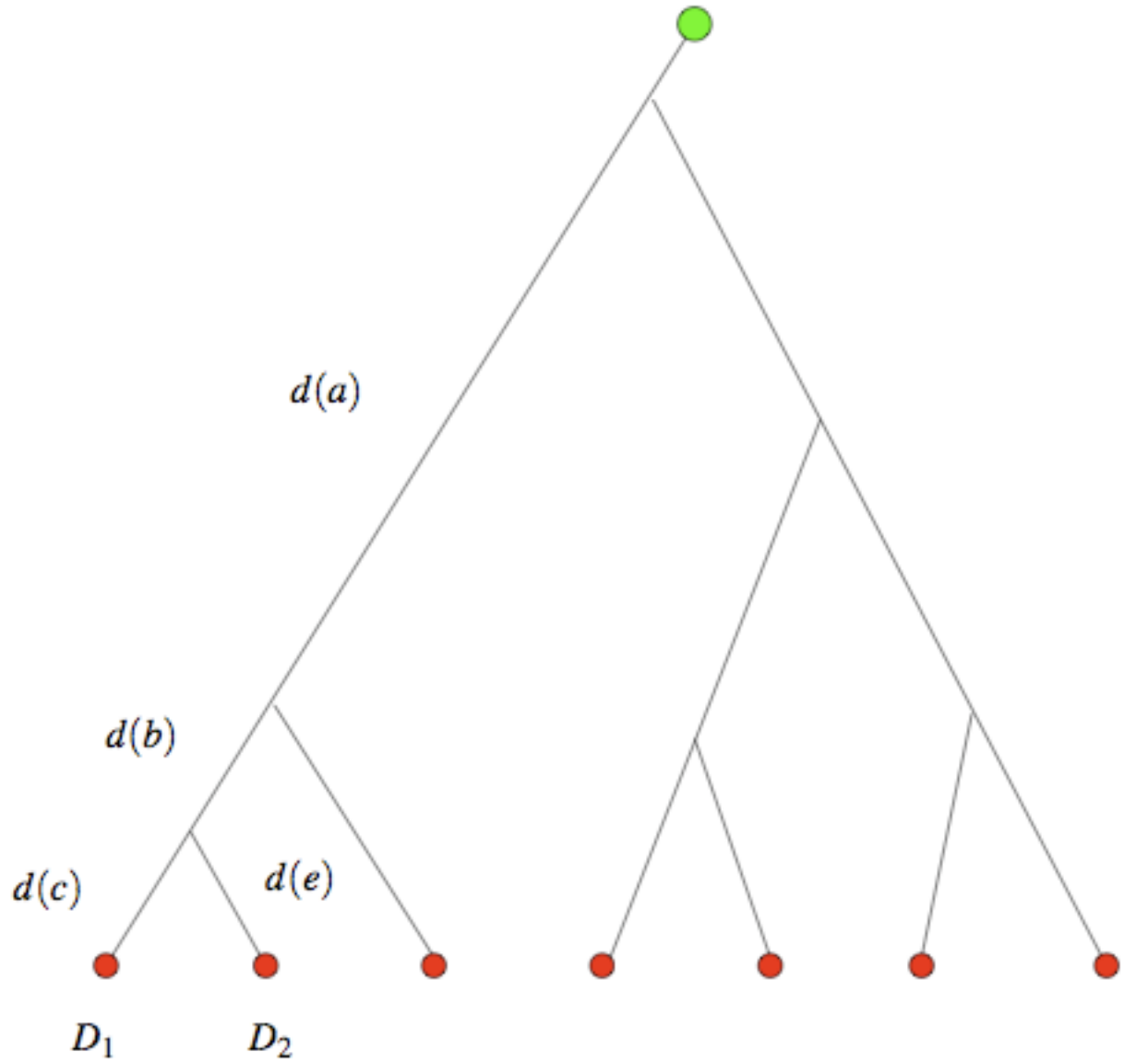


Figure 1: Multicast routing

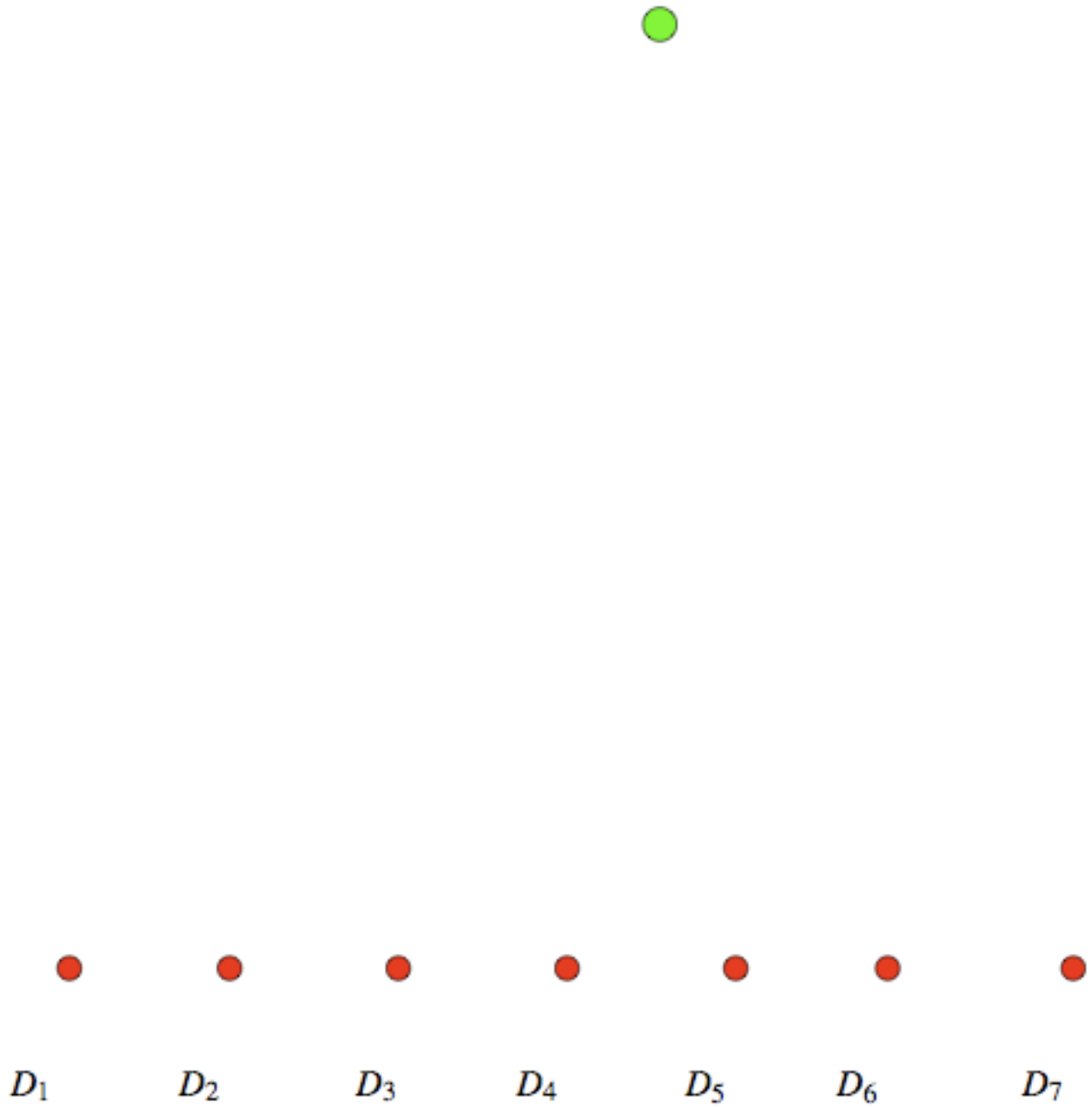


Figure 2: What we observe

Network Tomography

External Approach

Uses "'‘end-end"' measurements

Multicast Inference

- Single source allowed to send single data packet to some receivers via its routing tree.
- The packet duplicated at each branch point in the tree and sent further down the tree.

Questions

- Using Source to Destination delays can we recover the topology of the tree ?

Network Tomography

External Approach

Uses "'‘end-end"' measurements

Multicast Inference

- Single source allowed to send single data packet to some receivers via its routing tree.
- The packet duplicated at each branch point in the tree and sent further down the tree.

Questions

- Using Source to Destination delays can we recover the topology of the tree ?
- Again using the above delays can we understand the distribution of link delays of the tree ?

Internet Tomography

Basic Result

1. Assume that the tree is known, then using poly($\log n$) number of samples, we can estimate the delay distribution on edges with high accuracy.
2. Can estimate the unknown tree in $(\log n)^5$ samples.

Key ideas essentially from phylogenetics.

Edge flows on the complete graph

Problem setup

- Take complete graph. Attach exponential edge costs(lengths) IID with mean n.
- every node sends a flow of volume $1/n$ to every other node using the *least cost path*
- For ordered pair of nodes (s,t) let $\pi(s,t)$ denote the least cost path
- For any directed edge e the amount of flow passing through the edge is

$$F_n(e) = \frac{1}{n} \sum_{s,t} 1_{\{e \in \pi(s,t)\}}$$

Edge flows on the complete graph

Problem setup

- Take complete graph. Attach exponential edge costs(lengths) IID with mean n.
- every node sends a flow of volume $1/n$ to every other node using the *least cost path*
- For ordered pair of nodes (s,t) let $\pi(s,t)$ denote the least cost path
- For any directed edge e the amount of flow passing through the edge is

$$F_n(e) = \frac{1}{n} \sum_{s,t} 1_{\{e \in \pi(s,t)\}}$$

Math question:

Can we get asymptotics for the random empirical distribution of this edge flow namely

$$\frac{1}{n} \#\{e : F_n(e) > z \log n\} \tag{5}$$

Edge flows: Trying to prove

Result

As $n \to \infty$ for fixed $z > 0$,

$$\frac{1}{n}\#\{e : F_n(e) > z \log n\} \to_{L^1} G(z) \tag{6}$$

$G(z) = \int_0^\infty P(W_1 W_2 e^{-u} > z)\, du$
where W_1 and W_2 are independent Exponential(1).

Method of attack

Conceptual Idea

Fix a directed edge $e = (v_L, v_R)$ in the graph. If structure about edge e upto large distance known then value of flow through edge becomes *essentially* deterministic.

Method of attack

Conceptual Idea

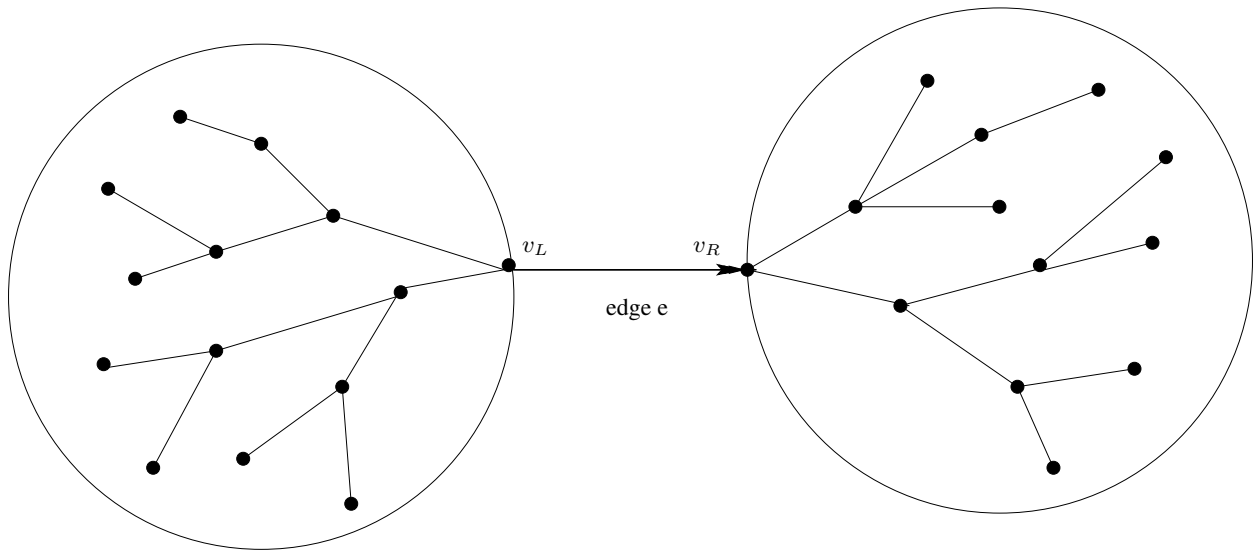
Fix a directed edge $e = (v_L, v_R)$ in the graph. If structure about edge e upto large distance known then value of flow through edge becomes *essentially* deterministic.

Math meaning:

Fix τ large

$$N_\tau = \{v : \min(D(v, v_L), D(v, v_R)) \leq \tau\}$$

with high prob. for large n and any fixed τ , N_τ is a tree



L(t) (vertices which are within τ from v_R)

R(t)(vertices which are within τ from v_R)

Figure 1: **Neighborhood about an edge**

Method of attack

Conceptual Idea

Fix a directed edge $e = (v_L, v_R)$ in the graph. If structure about edge e upto large distance known then value of flow through edge becomes *essentially* deterministic.

Math meaning:

Fix τ large

$$N_\tau = \{v : \min(D(v, v_L), D(v, v_R)) \leq \tau\}$$

with high prob. for large n and any fixed τ , N_τ is a tree
Suppose we show

$$E(F_n(e)|N_\tau) \approx \frac{\#R(\tau)}{e^\tau} . \frac{\#L(\tau)}{e^\tau} e^{-L(e)} \log n$$

$$Var(F_n(e)/\log n|N_\tau) \approx 0 \text{ for large n and large } \tau$$

Method of Attack contd.

Above $\Longrightarrow$

$$P\left(\text{nbhd a tree } |\frac{F_n(e)}{\log n} - \frac{\#R(\tau)}{e^{\tau}}.\frac{\#L(\tau)}{e^{\tau}}e^{-L(e)}| > \epsilon | L(e)\right) \approx 1$$

for large n and large τ

For large n

$$\#R(\tau), \#L(\tau)$$

are basically independant with the same distribution as $N(\tau)$ the number of offspring in a Yule process observed for time τ

Method of Attack contd.

Above $\Longrightarrow$

$$P\left(\text{nbhd a tree } |\frac{F_n(e)}{\log n} - \frac{\#R(\tau)}{e^\tau}.\frac{\#L(\tau)}{e^\tau}e^{-L(e)}| > \epsilon | L(e)\right) \approx 1$$

for large n and large τ

For large n

$$\#R(\tau), \#L(\tau)$$

are basically independant with the same distribution as $N(\tau)$ the number of offspring in a Yule process observed for time τ

Behavior of functionals

- For large τ ; $N(\tau)/e^\tau \approx W$ where W is an $\exp(1)$. Thus $\#R(\tau)/e^\tau \sim W_1, \#L(\tau)/e^\tau \sim W_2$
- The empirical distribution

$$\frac{1}{n}\{\#e : L_e \in [a,b]\} \Longrightarrow_v \lambda[a,b]$$

Method of attack contd.

Heuristically

"Empirical Distribution" of normalized flow volumes $\frac{F_n(e)}{\log n} \approx W_1 W_2 e^{-L}$ where $L \sim \lambda$

Outline of the expectation result

- Want to show :

$$\mathbb{E}(F_n(e)|N_\tau) \sim \frac{\#R(\tau)}{e^\tau} \frac{\#R(\tau)}{e^\tau} \log n$$

- By symmetry

$$\mathbb{E}(F_n(e)) = \mathbb{E}(M_1)$$

$$M_1 := \{\#s : e \in \pi(1, s)\}$$

First passage percolation from node 1 [$E(M_1)$]

Flow from node 1

- Start a flow moving at rate 1 from node 1.
- Let $S_{n,k}$ be the time to see the k^{th} distinct vertex, $N_n(t)$ number of nodes at distance t from vertex 1.
- Distribution

$$(S_{n,k+1} - S_{n,k}) \stackrel{d}{=} (\frac{n}{k(n-k)} Y_i)$$

- Yule process: Continuous time birth process where each individual produces new offspring at rate 1. For S_k, $N(t)$

$$(S_{k+1} - S_k)_{i \geq 1} \stackrel{d}{=} (\frac{1}{k} Y_i)_{i \geq 1}$$

- $N(t) \sim We^t$.

Computing $E(M_1)$

Polya urn type computation implies

$$\frac{1}{n}\mathbb{E}(M_1) \approx \mathbb{E}(\frac{\#R(\tau)}{N_n(T+\sigma)})$$

$T :=$ time to hit nbhd, $\sigma = 2\tau + L(e)$, the diameter of the neighborhood.

Now

$$\mathbb{E}(\frac{\#R(\tau)}{N_n(T+\sigma)}) \approx \#R(\tau)\sum_k \frac{\#L(\tau)}{n}\mathbb{E}(\frac{1}{N_n(T+\sigma)}|N_n(T)=k)$$

Use the fact that uniformly $N_n(T+\sigma) \sim N_n(T)e^{\sigma}$ to get

$$\frac{1}{n}\mathbb{E}(M_1) \approx \frac{\#L(\tau)}{e^{\tau}}\frac{\#R(\tau)}{e^{\tau}}\frac{1}{n}\sum_1^n \frac{1}{i}$$

Predictions for other models

number of models where asymptotics for degree distribution known i.e $\exists$ a distribution on the set of integers such that:

$$P(D_\infty = i) = d_i(\text{ e.g. } = \frac{C}{k^\tau})$$

- Attach exp(1) edge lengths.
- structure of the graph looking outwards from typical node looks like a continuous time branching process with offspring distribution

$$P(D^* = i) = (i+1)P(D_\infty = i+1)/E(D_\infty)$$

Predictions for other models contd.

$\exists$ random variable Z and a Malthusian growth parameter θ such that $\frac{N_\infty(t)}{\exp(\theta t)} \to Z$.

Predictions for other models contd.

$\exists$ random variable Z and a Malthusian growth parameter θ such that $\frac{N_\infty(t)}{\exp(\theta t)} \to Z$.

The above argument for complete graph suggests

$$\frac{1}{n}\#\{e : F_n(e) > z \log n\} \to G(z)$$

$G(z) = \int_0^\infty P(Z_1 Z_2 e^{-\theta u} > z)\, d\mu(u)$
i.e distribution of $Z_1 Z_2 \exp(-\theta\xi)$, $\xi \sim \exp(1)$

Predictions for other models contd.

$\exists$ random variable Z and a Malthusian growth parameter θ such that $\frac{N_\infty(t)}{\exp(\theta t)} \to Z$.

The above argument for complete graph suggests

$$\frac{1}{n}\#\{e : F_n(e) > z \log n\} \to G(z)$$

$G(z) = \int_0^\infty P(Z_1 Z_2 e^{-\theta u} > z)\, d\mu(u)$
i.e distribution of $Z_1 Z_2 \exp(-\theta\xi)$, $\xi \sim \exp(1)$

Special case: Random r regular graph

$Z \sim$ Gamma$(r-2)$, $\theta = 1$

Predictions for other models contd

Multiple source destination pairs

- Complete graph with exponential (1) edge costs
- Fix k nodes and let D_{ij} be the least cost path between them.
- Then

$$(nD_{ij} - \log n : i \neq j) \Longrightarrow (\xi_i + \xi_j - \xi_{ij})$$

where $\xi_i \sim$ double exponential distribution

ultra-small world to *small world*

- wide array of real world networks with degree exponent in $[2, 3]$.
- graph distance for corresponding rand. graph model $\ll \log n$ e.g $\log \log n$
- possible to prove that in these models attaching random edge weights (congestion costs) causes the following behavior

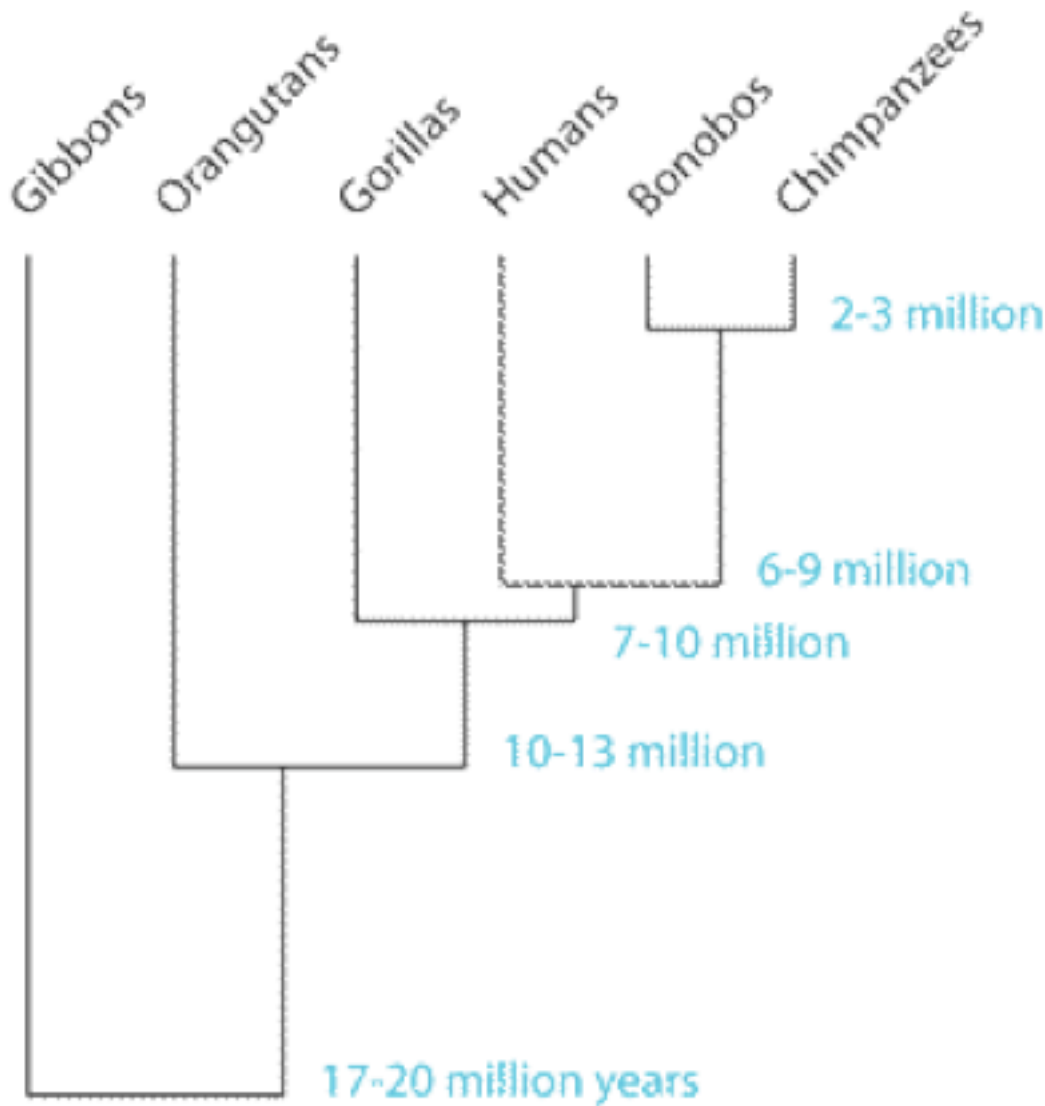
$$\frac{H_n}{\log n} \longrightarrow C$$

Phylogenetics

Main Problem

- Infer evolutionary histories from molecular data.
- Evolution represented as a tree where branching points indicate speciation events
- Molecular data assumed to evolve according to some kind of a Markov model on trees
- Measurement of sequences at leaves (extant species) try to reconstruct the tree.

hominoid_mol_phylogeny.png (PNG Image, 300x338 pixels)



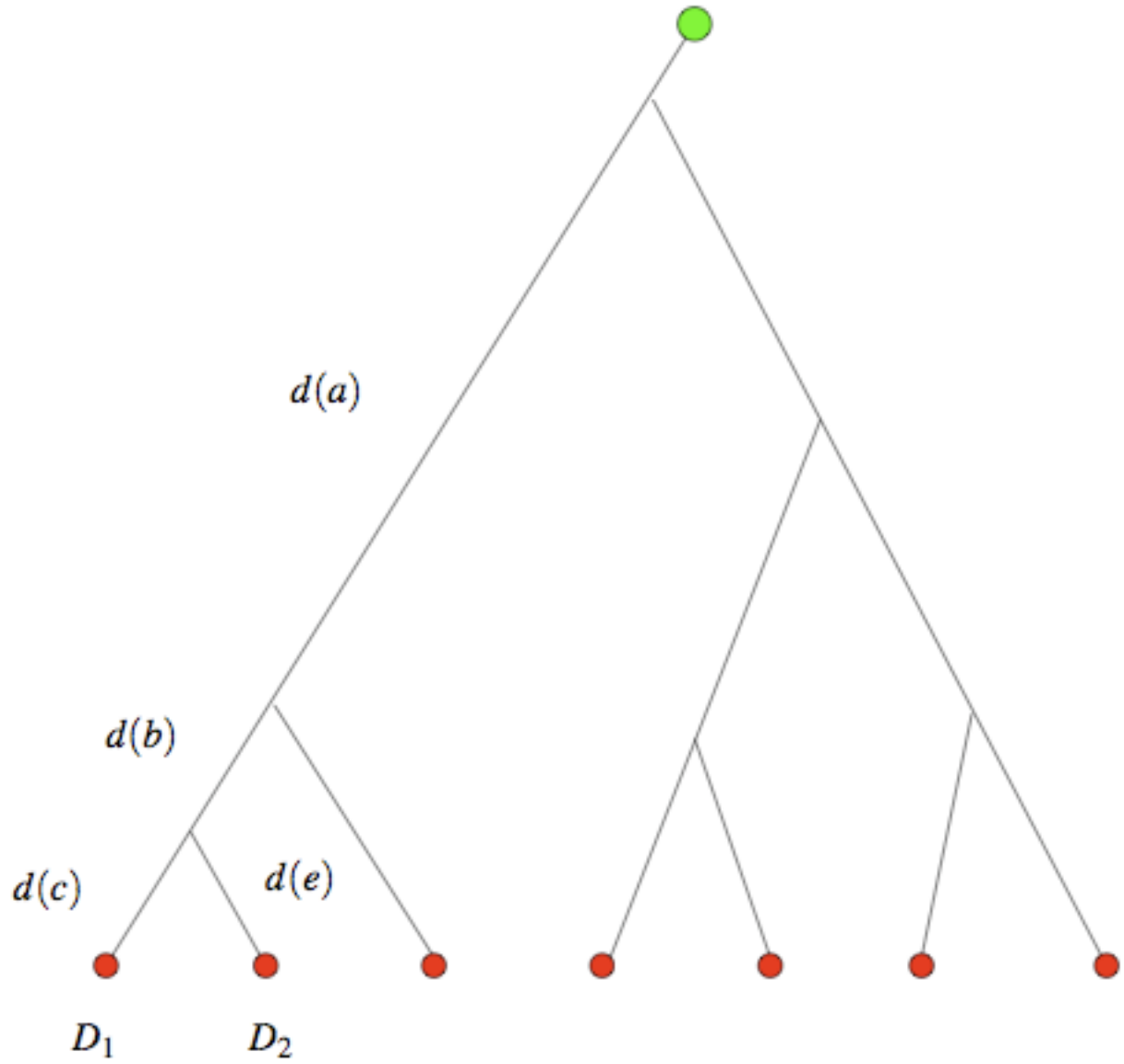


Figure 1: Multicast routing

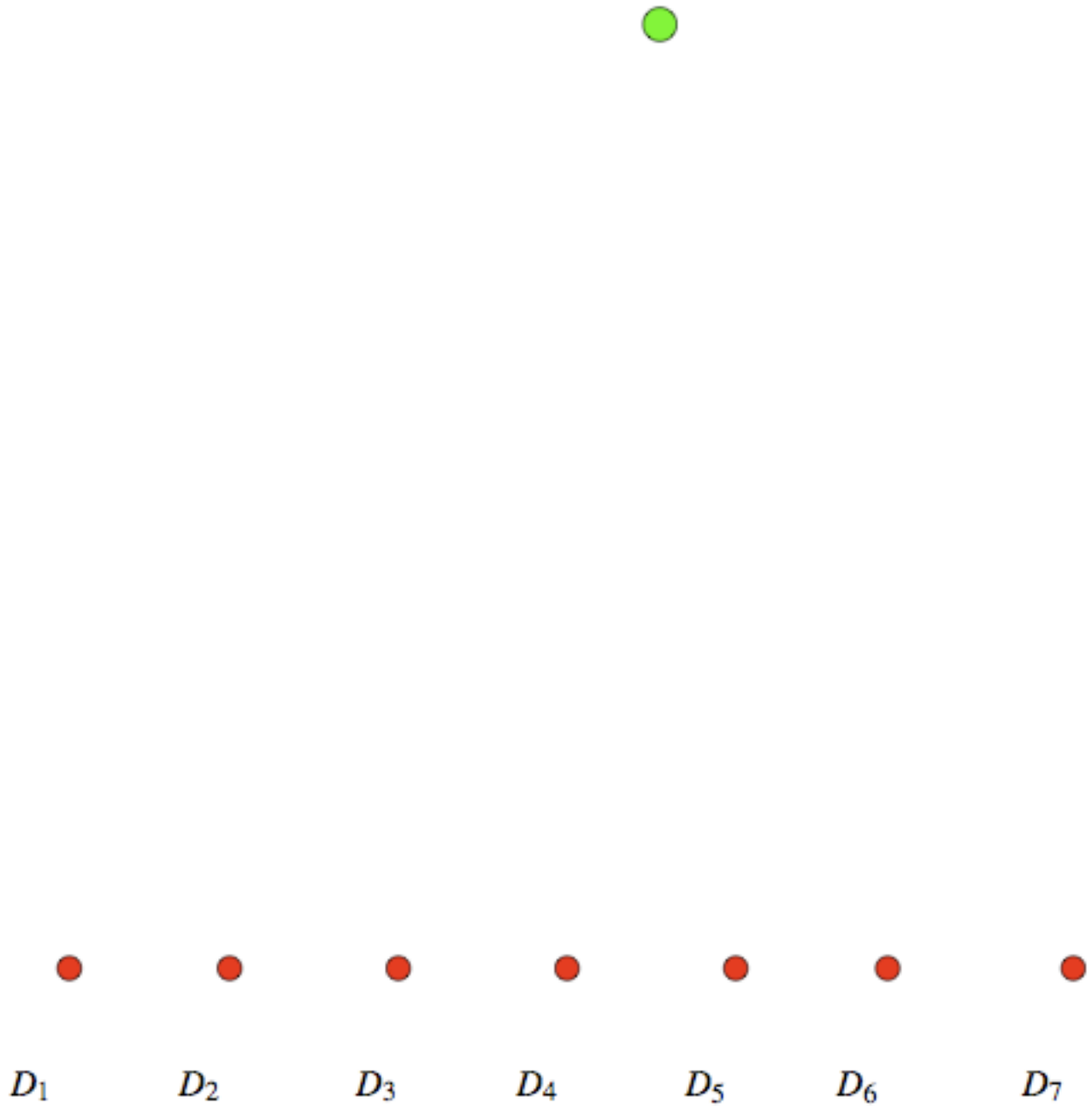


Figure 2: What we observe

Phylogenetics

Main Problem

- Infer evolutionary histories from molecular data.
- Evolution represented as a tree where branching points indicate speciation events
- Molecular data assumed to evolve according to some kind of a Markov model on trees
- Measurement of sequences at leaves (extant species) try to reconstruct the tree.
- Methods of reconstruction: Maximum Likelihood, MCMC, distance-based methods

Tree metrics

For 2 receivers:

$$\begin{aligned} W^{(2)}(a,b) &= Var(D_a - D_b) \\ &= \sum_{e \in P_{a,b}} w_e^{(2)} \end{aligned}$$

$w_e^{(2)} = \mathrm{Var}(d_e)$is an example of a "‘tree metric"’ **Assumption:** $w_e^{(2)} > f \ \forall e$

Tree metrics

For 2 receivers:

$$\begin{aligned} W^{(2)}(a,b) &= Var(D_a - D_b) \\ &= \sum_{e \in P_{a,b}} w_e^{(2)} \end{aligned}$$

$w_e^{(2)} = \mathrm{Var}(d_e)$is an example of a "'tree metric"' **Assumption:** $w_e^{(2)} > f \ \forall e$

Use of tree metrics : FPM

A non degenerate tree metric helps us to determine all quartets, i.e. $ab|cd$ iff

$$W(a,b) + W(c,d) \leq \min\{W(a,c) + W(b,d), W(a,d) + W(b,c)\}.$$

Tree metrics

For 2 receivers:

$$\begin{aligned} W^{(2)}(a,b) &= Var(D_a - D_b) \\ &= \sum_{e \in P_{a,b}} w_e^{(2)} \end{aligned}$$

$w_e^{(2)} = \mathrm{Var}(d_e)$is an example of a "'tree metric"' **Assumption:** $w_e^{(2)} > f \ \forall e$

Use of tree metrics : FPM

A non degenerate tree metric helps us to determine all quartets, i.e. $ab|cd$ iff

$$W(a,b) + W(c,d) \leq \min\{W(a,c) + W(b,d), W(a,d) + W(b,c)\}.$$

Once all quartets identified then entire tree can be reconstructed.

- So idea: Estimate $W^{(2)}(a, b)$ accurately and then done.

- So idea: Estimate $W^{(2)}(a, b)$ accurately and then done.
- Problem: Caterpillar trees (Diam of typical binary tree = 0($\sqrt{n}$)) Would cause too much congestion in network.

- So idea: Estimate $W^{(2)}(a, b)$ accurately and then done.
- Problem: Caterpillar trees (Diam of typical binary tree = 0($\sqrt{n}$)) Would cause too much congestion in network.
- Can we do better ? Can we use small paths ?

- So idea: Estimate $W^{(2)}(a,b)$ accurately and then done.
- Problem: Caterpillar trees (Diam of typical binary tree = 0($\sqrt{n}$)) Would cause too much congestion in network.
- Can we do better ? Can we use small paths ?

Depth of a tree

$e = (u,v)$ be an edge. $\gamma_u(e)$ the shortest distance from u to the leaves not using edge e. depth of T

$$\Delta(T) = \max_{e=(u,v)\in E} \max\{\gamma_u(e), \gamma_v(e)\}.$$

- So idea: Estimate $W^{(2)}(a,b)$ accurately and then done.
- Problem: Caterpillar trees (Diam of typical binary tree = 0($\sqrt{n}$)) Would cause too much congestion in network.
- Can we do better ? Can we use small paths ?

Depth of a tree

$e = (u,v)$ be an edge. $\gamma_u(e)$ the shortest distance from u to the leaves not using edge e. depth of T

$$\Delta(T) = \max_{e=(u,v)\in E} \max\{\gamma_u(e), \gamma_v(e)\}.$$

Point

$$\Delta(T) \leq \log n$$

and typically is of order $\log_2 \log_2 n$

Can we use only short paths ?

Short quartet

Quartet $\{a, b, c, d\}$ called short if for all $i, j \in \{a, b, c, d\}$, we have

$$\text{graph dist} \leq (2\,\Delta(T) + 3)$$

Can we use only short paths ?

Short quartet

Quartet $\{a, b, c, d\}$ called short if for all $i, j \in \{a, b, c, d\}$, we have

$$\text{graph dist} \leq (2\,\Delta(T) + 3)$$

Combinatorial Fact [ESSW]

If we have all short quartets than using rules like

$$\text{if } ab|cd \text{ and } ab|ce \text{ are quartet splits of } T, \text{ then so is } ab|de.$$

we can reconstruct the entire tree.

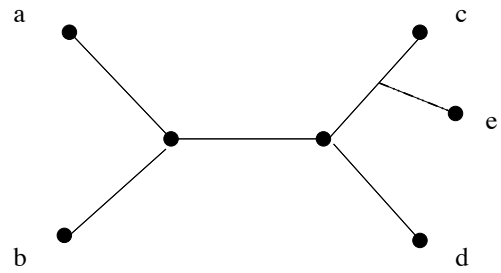
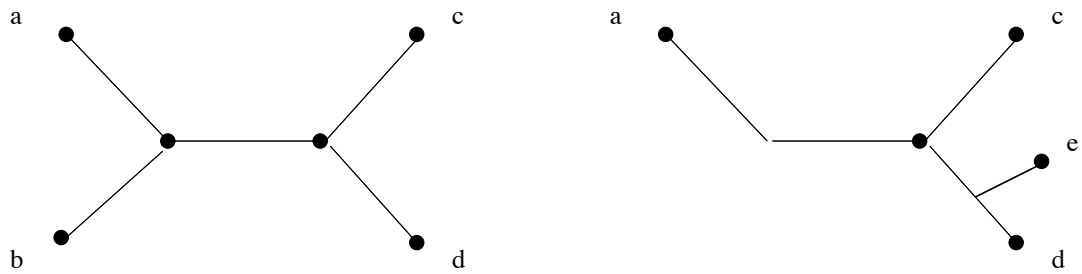


Figure 1: Dyadic Rule 1: $ab|cd$ and $ab|ce \Rightarrow ab|ce$



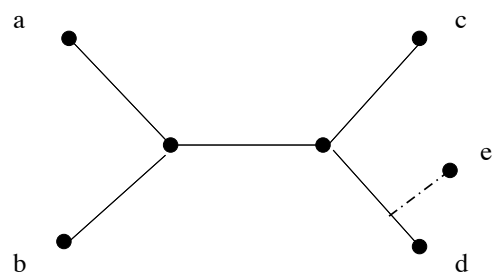


Figure 2: Dyadic Rule 2: $ab|cd$ and $ac|de \Rightarrow bc|de, ab|ce, ab|de$

Can we use only short paths ?

Short quartet

Quartet $\{a, b, c, d\}$ called short if for all $i, j \in \{a, b, c, d\}$, we have

$$\text{graph dist} \leq (2\,\Delta(T) + 3)$$

Combinatorial Fact [ESSW]

If we have all short quartets than using rules like

$$\text{if } ab|cd \text{ and } ab|ce \text{ are quartet splits of } T, \text{ then so is } ab|de.$$

we can reconstruct the entire tree.

Approximate metrics

Four-Point Method applied to $\hat{W}$ infers the right quartet split for T if the following condition holds. For all $i, j \in \{a, b, c, d\}$, we have

$$\left|W(i,j) - \hat{W}(i,j)\right| < \frac{f}{2}.$$

Basic Algorithm for Reconstructing the tree

1. Estimate $\mathrm{Var}(D_a - D_b)$ by the sample estimate. Thus get $\hat{W}^{(2)}(a, b)$

Basic Algorithm for Reconstructing the tree

1. Estimate $\mathrm{Var}(D_a - D_b)$ by the sample estimate. Thus get $\hat{W}^{(2)}(a, b)$
2. Call $q = \{a, b, c, d\}$ an *estimated short quartet* if for all $i, j \in \{a, b, c, d\}$, we have

$$\hat{W}^{(2)}(i, j) \leq (2 \log n + 3)M + \frac{M}{2}. \tag{7}$$

Basic Algorithm for Reconstructing the tree

1. Estimate $\mathrm{Var}(D_a - D_b)$ by the sample estimate. Thus get $\hat{W}^{(2)}(a,b)$
2. Call $q = \{a,b,c,d\}$ an *estimated short quartet* if for all $i,j \in \{a,b,c,d\}$, we have

$$\hat{W}^{(2)}(i,j) \leq (2\log n + 3)M + \frac{M}{2}. \tag{7}$$

3. Get all estimated short quartets from data.

Basic Algorithm for Reconstructing the tree

1. Estimate $\mathrm{Var}(D_a - D_b)$ by the sample estimate. Thus get $\hat{W}^{(2)}(a,b)$
2. Call $q = \{a,b,c,d\}$ an *estimated short quartet* if for all $i,j \in \{a,b,c,d\}$, we have

$$\hat{W}^{(2)}(i,j) \leq (2\log n + 3)M + \frac{M}{2}. \tag{7}$$

3. Get all estimated short quartets from data.
4. Get the Dyadic Closure of above quartets.

Basic Algorithm for Reconstructing the tree

1. Estimate $\mathrm{Var}(D_a - D_b)$ by the sample estimate. Thus get $\hat{W}^{(2)}(a,b)$
2. Call $q = \{a,b,c,d\}$ an *estimated short quartet* if for all $i,j \in \{a,b,c,d\}$, we have

$$\hat{W}^{(2)}(i,j) \leq (2\log n + 3)M + \frac{M}{2}. \tag{7}$$

3. Get all estimated short quartets from data.
4. Get the Dyadic Closure of above quartets.
5. Get the entire tree.

Basic Result

If we take $k = (\log n)^5$ samples then enough to reconstruct the tree with high accuracy

Conclusion

- Lots of interesting open problems.
  - Vickrey Clarke Grooves measure of overpayment and shortest path trees
  - Gossip algorithms and first passage percolation
  - Viral marketing and Statistical physics
  - Phase transitions and Tomography
  - Constructing optimal networks

Conclusion

- Lots of interesting open problems.
  - Vickrey Clarke Grooves measure of overpayment and shortest path trees
  - Gossip algorithms and first passage percolation
  - Viral marketing and Statistical physics
  - Phase transitions and Tomography
  - Constructing optimal networks
- We as probabilists have an advantage !! Refined tools like branching process asymptotics, martingales, large deviations, percolation, come in very useful.