Trace Complexity

Flavio Chierichetti
Sapienza University of Rome

Online social networks

The total number of
active user accounts
exceeds
two billions!

Advertisement

Online Social Networks

Sociology

Economics

Political Science

Computer Science

Online Social Networks

Sociology
Economics
Political Science

Computer Science

Models of human behavior

Models Improvements ← Models Validations

Online Social Networks

Sociology

Economics ⟷ Computer Science

Political Science

Models of human behavior

Models Improvements ← **Models Validations**

Empirical Analysis

- This sort of analysis is becoming more and more important in social networks research
- Exact user data is **very hard** to obtain for **privacy** related reasons

Empirical Analysis

- This sort of analysis is becoming more and more important in social networks research
- Exact user data is **very hard** to obtain for **privacy** related reasons
- Empirical researchers have a hard time testing their hypotheses / validating their models

“Data Traces”

Researchers try to **guess**
the *missing data*,
using *existing observations* and a
reasonable *theoretical model*

“Data Traces”

Researchers try to **guess**
the *missing data*,
using *existing observations* and a
reasonable *theoretical model*

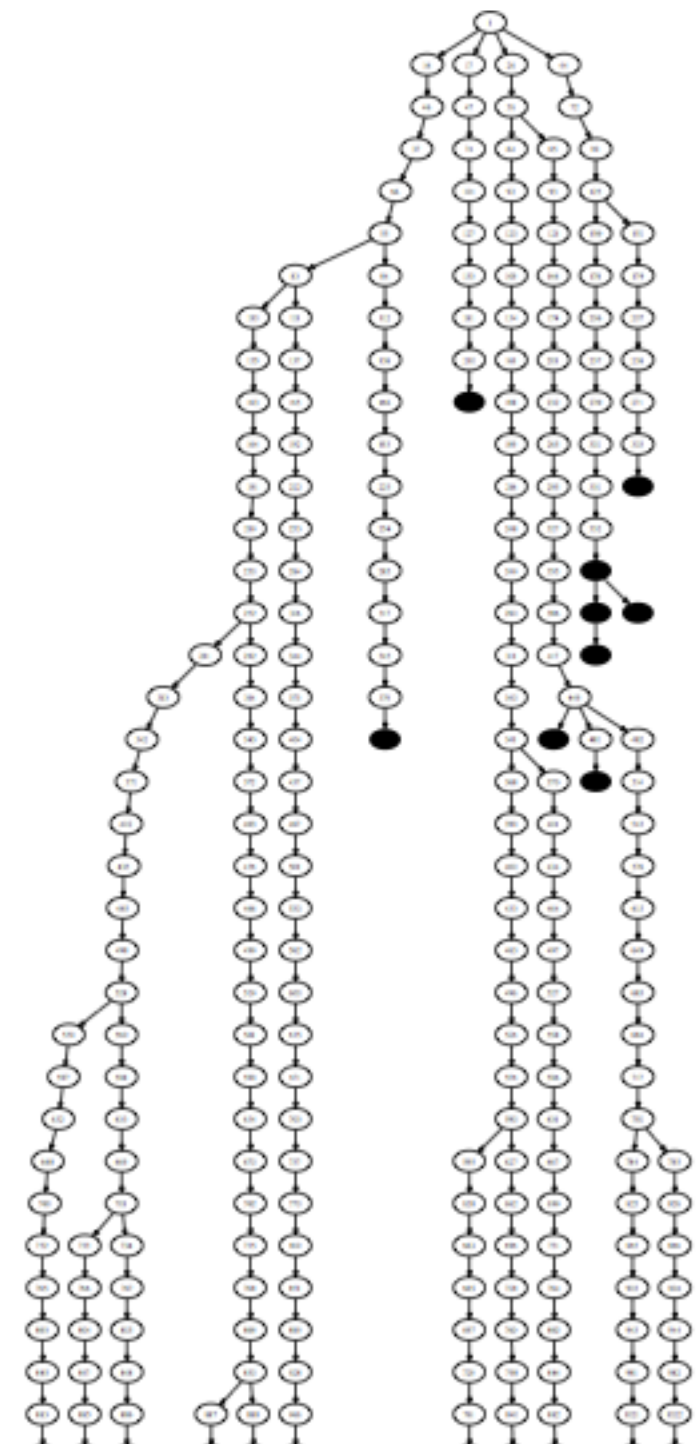
Theoretical researchers can help by
trying to understand whether
sound guessing algorithms *exist*

Online Guessing Tasks

- Which of these people are friends?

Online Guessing Tasks

- Which of these people are friends?
- How many people were infected by this meme?



Online Guessing Tasks

- We see little pieces of some online social process.
- We would like to make some **sensible** guesses on the process as a whole.

Network Reconstruction

Joint with

Bruno Abrahao, Robert Kleinberg, Alessandro Panconesi

Network Reconstruction

- Incomplete Traces of Information Flow.
- Wish to infer the hidden network.

Network Reconstruction

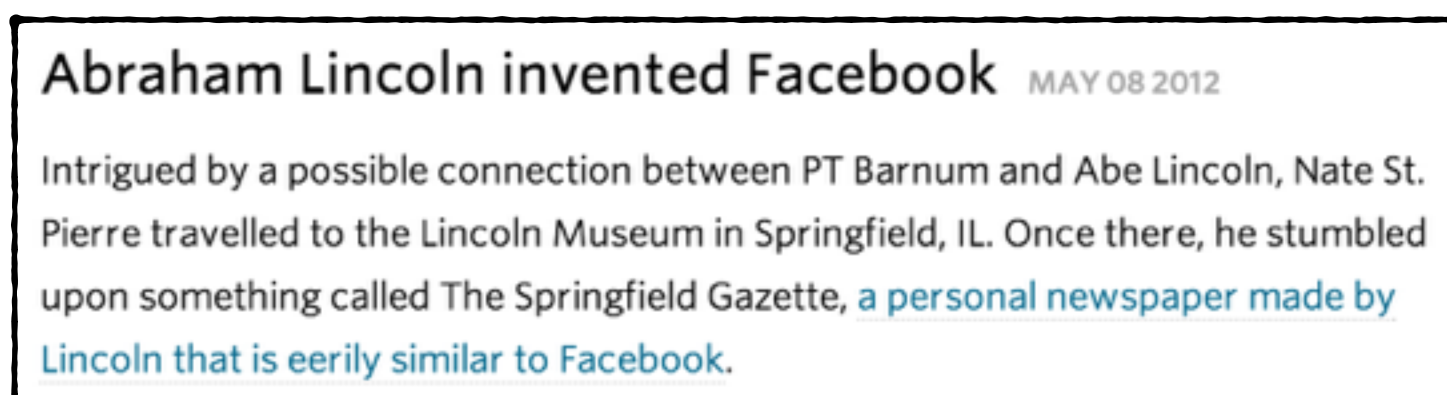
- Rich area of study, pioneered by [Adar, Adamic, ’05] in the context of social networks.
- MLE-based approaches:
  - [Gomez-Rodriguez, Leskovec, Krause ’10],
  - [Gomez-Rodriguez, Balduzzi, Scholkopf ’11],
  - [Myers, Leskovec ’11],
  - [Du et al. ’12]
- Information theoretic approaches
  - [Netrapalli, Sanghavi ’12],
  - [Grippon, Rabbat ’13]

“Inferring networks of diffusion and influence”

Gomez-Rodriguez, Leskovec, Krause *[KDD’10]*

“

- Gomez-Rodriguez et al studied a large collection of blogs and memes to guess the *blogger network* that allowed *memes* to *spread*.


Abraham Lincoln invented Facebook MAY 08 2012

Intrigued by a possible connection between PT Barnum and Abe Lincoln, Nate St. Pierre travelled to the Lincoln Museum in Springfield, IL. Once there, he stumbled upon something called The Springfield Gazette, a personal newspaper made by Lincoln that is eerily similar to Facebook.

“Inferring networks of diffusion and influence”

Gomez-Rodriguez, Leskovec, Krause *[KDD’10]*

“

- Gomez-Rodriguez et al studied a large collection of blogs and memes to guess the *blogger network* that allowed *memes* to *spread*.

- They proposed a random meme-diffusion model, and used it for making the guess.

Flow of Memes

Alice's Blog

Bob's Blog - Live from Lewisville

c.
charlie's blog

The World According to Dave
My unsolicited ramblings, rants, musings & bloviations (totally free and worth every penny)

Flow of Memes

Alice's Blog

Bob's Blog - Live from Lewisville

Mar 4, 2014, 8:25am
Mrs. Hudson's cake shop will reopen!

c.
charlie's blog

The World According to Dave
My unsolicited ramblings, rants, musings & bloviations (totally free and worth every penny)

Flow of Memes

Alice's Blog

Mar 4, 2014, 9:00am
Rejoice! Mrs. Hudson cake shop is reopening

Bob's Blog - Live from Lewisville

Mar 4, 2014, 8:25am
Mrs. Hudson's cake shop will reopen!

c.
charlie's blog

The World According to Dave
My unsolicited ramblings, rants, musings & bloviations (totally free and worth every penny)

Flow of Memes

Alice's Blog

Mar 4, 2014, 9:00am
Rejoice! Mrs. Hudson cake shop is reopening

Bob's Blog - Live from Lewisville

Mar 4, 2014, 8:25am
Mrs. Hudson's cake shop will reopen!

c.
charlie's blog

Mar 4, 2014, 10:00am
OMG! Hudson's is back in business!

The World According to Dave
My unsolicited ramblings, rants, musings & bloviations (totally free and worth every penny)

Flow of Memes

Alice's Blog

Mar 4, 2014, 9:00am
Rejoice! Mrs. Hudson cake shop is reopening

Bob's Blog - Live from Lewisville

Mar 4, 2014, 8:25am
Mrs. Hudson's cake shop will reopen!

c.
charlie's blog

Mar 4, 2014, 10:00am
OMG! Hudson's is back in business!

The World According to Dave
My unsolicited ramblings, rants, musings & bloviations (totally free and worth every penny)

Mar 4, 2014, 10:30am
Hudson's is about to reopen! Slurp!

Flow of Memes

Alice's Blog

Mar 4, 2014, 9:00am
Rejoice! Mrs. Hudson cake shop is reopening

Bob's Blog - Live from Lewisville

Mar 4, 2014, 8:25am
Mrs. Hudson's cake shop will reopen!

c.
charlie's blog

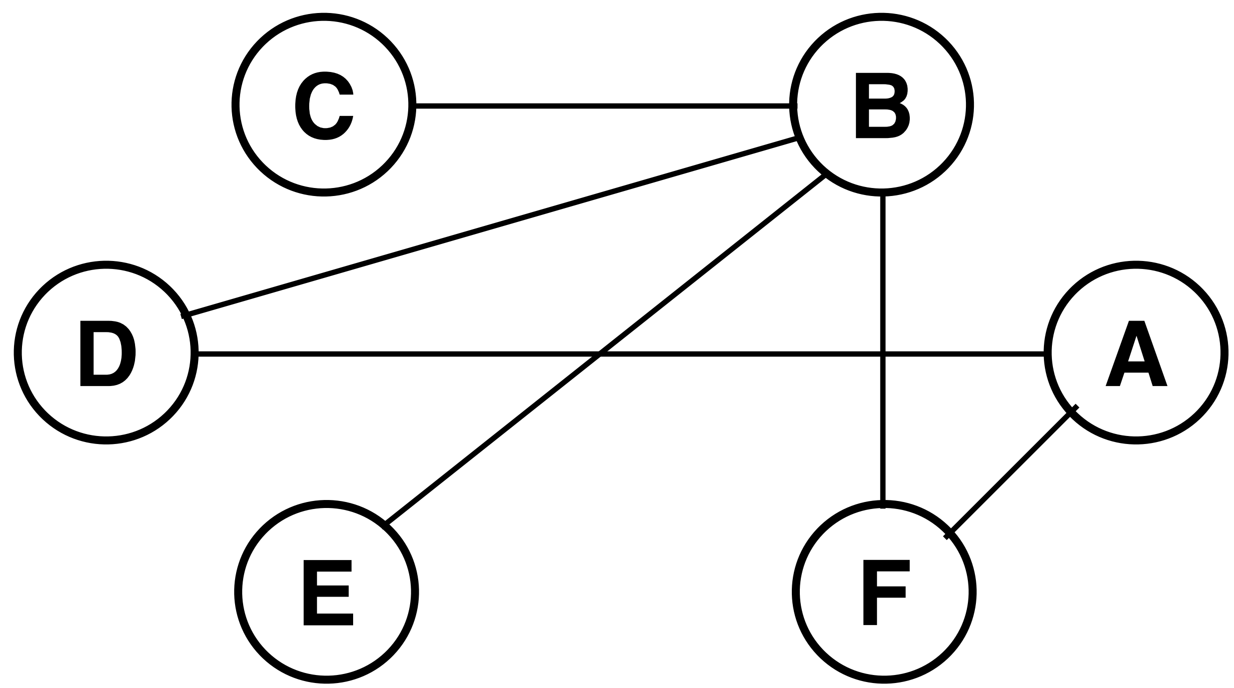
Mar 4, 2014, 10:00am
OMG! Hudson's is back in business!

The World According to Dave
My unsolicited ramblings, rants, musings & bloviations (totally free and worth every penny)

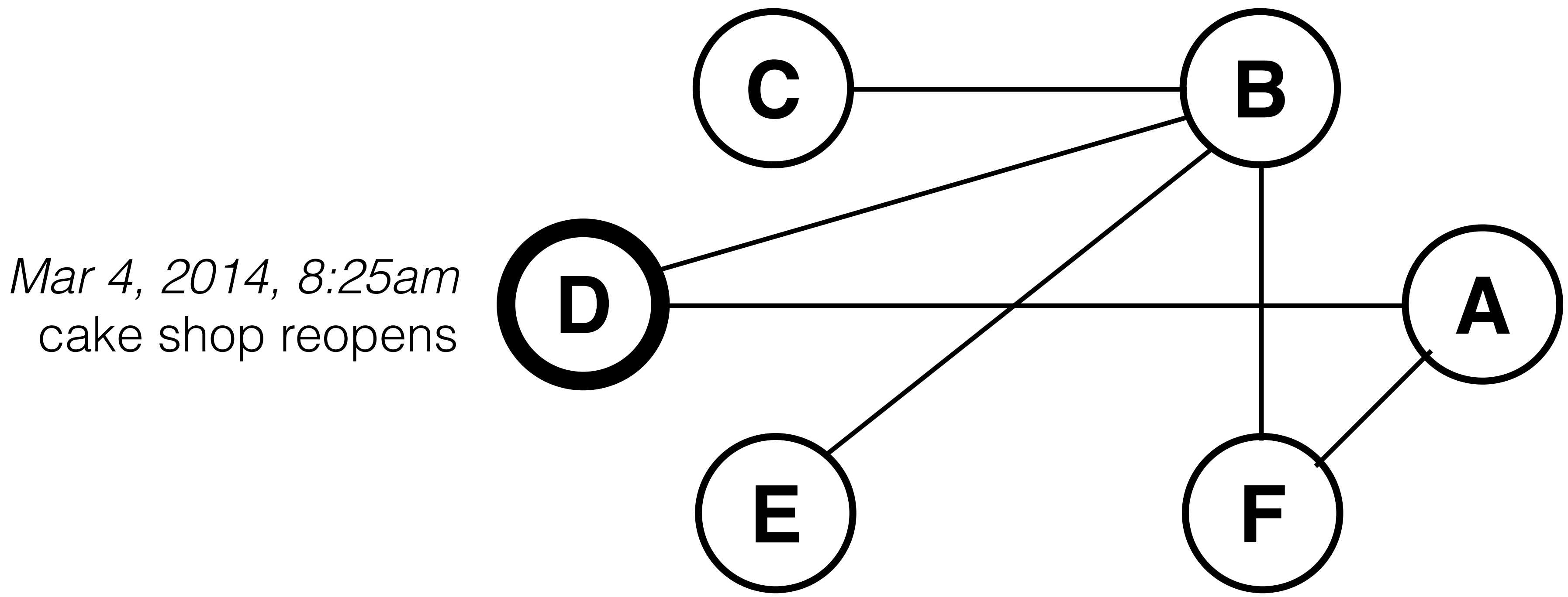
Mar 4, 2014, 10:30am
Hudson's is about to reopen! Slurp!

Can we infer which blogger follows which blog?

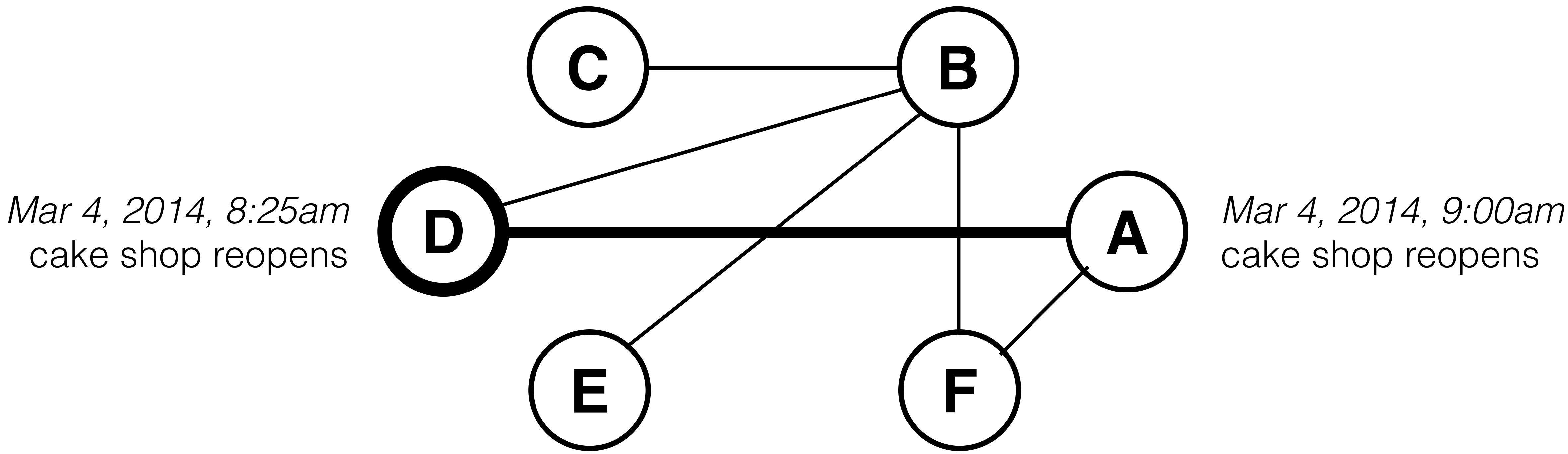
Meme Flow



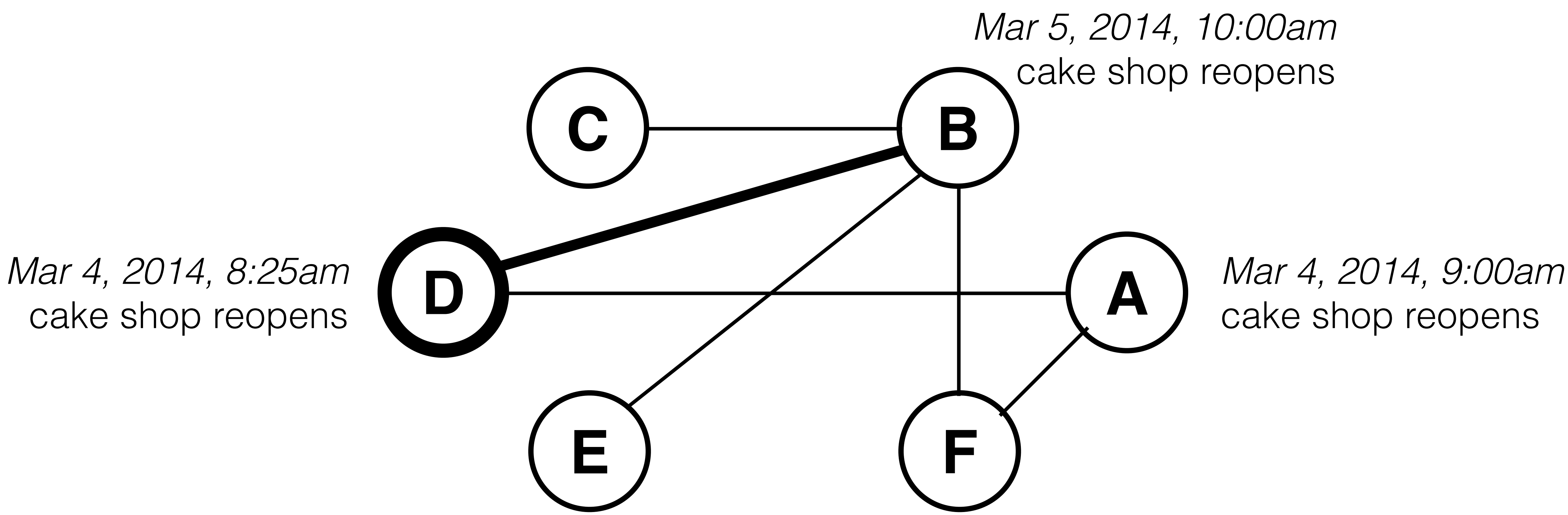
Meme Flow



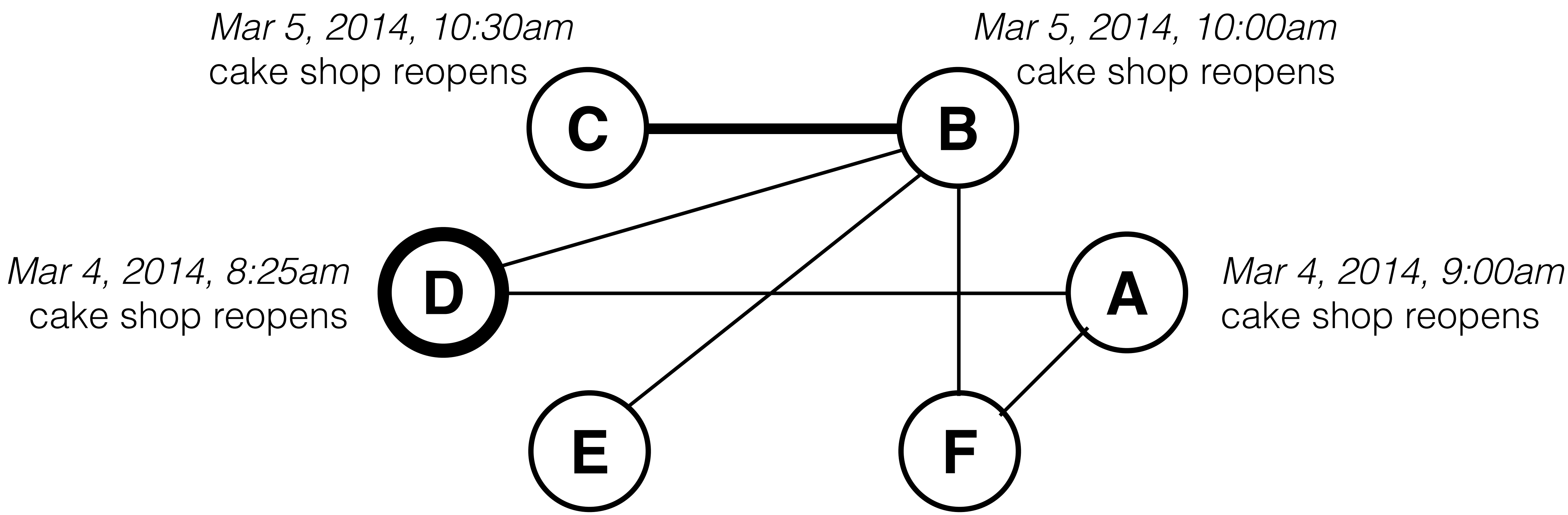
Meme Flow


C
B
D
A
E
F
Mar 4, 2014, 8:25am
cake shop reopens
Mar 4, 2014, 9:00am
cake shop reopens

Meme Flow

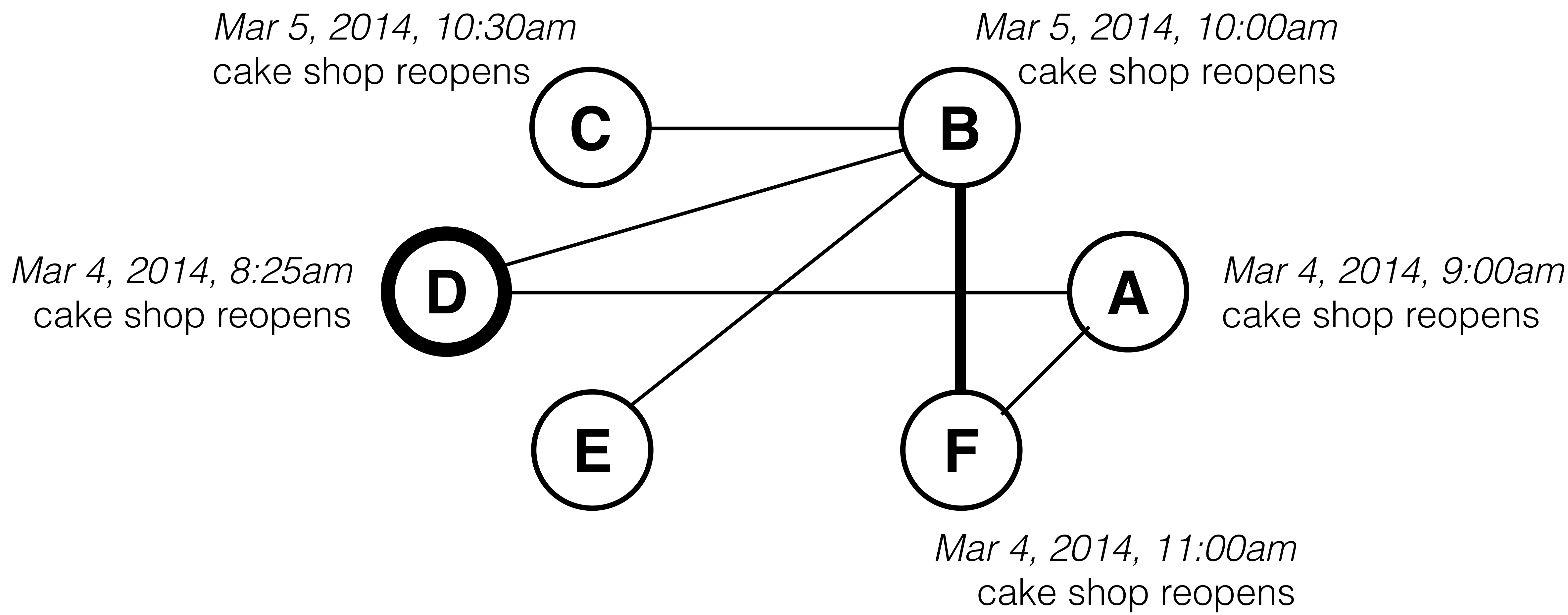

Mar 5, 2014, 10:00am
cake shop reopens
C
B
Mar 4, 2014, 8:25am
cake shop reopens
D
A
Mar 4, 2014, 9:00am
cake shop reopens
E
F

Meme Flow



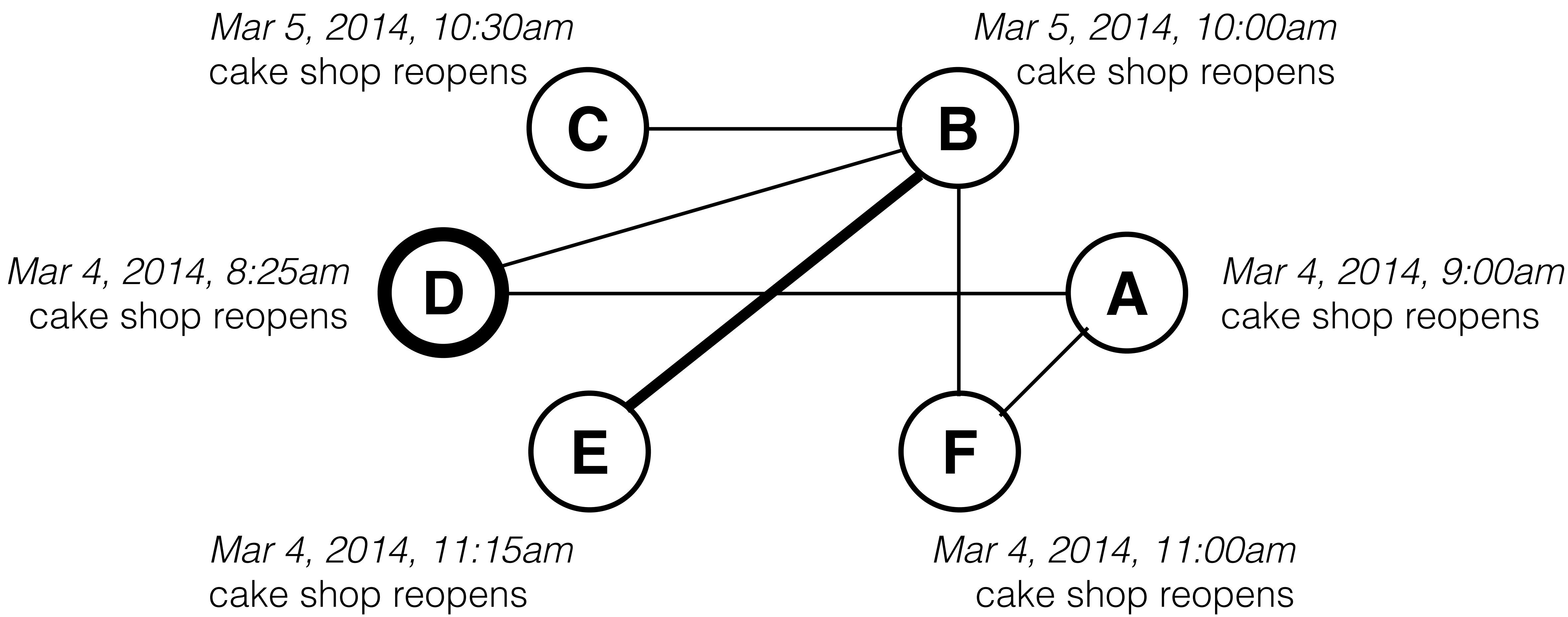
Mar 5, 2014, 10:30am
cake shop reopens
C
B
Mar 5, 2014, 10:00am
cake shop reopens
Mar 4, 2014, 8:25am
cake shop reopens
D
A
Mar 4, 2014, 9:00am
cake shop reopens
E
F

Meme Flow

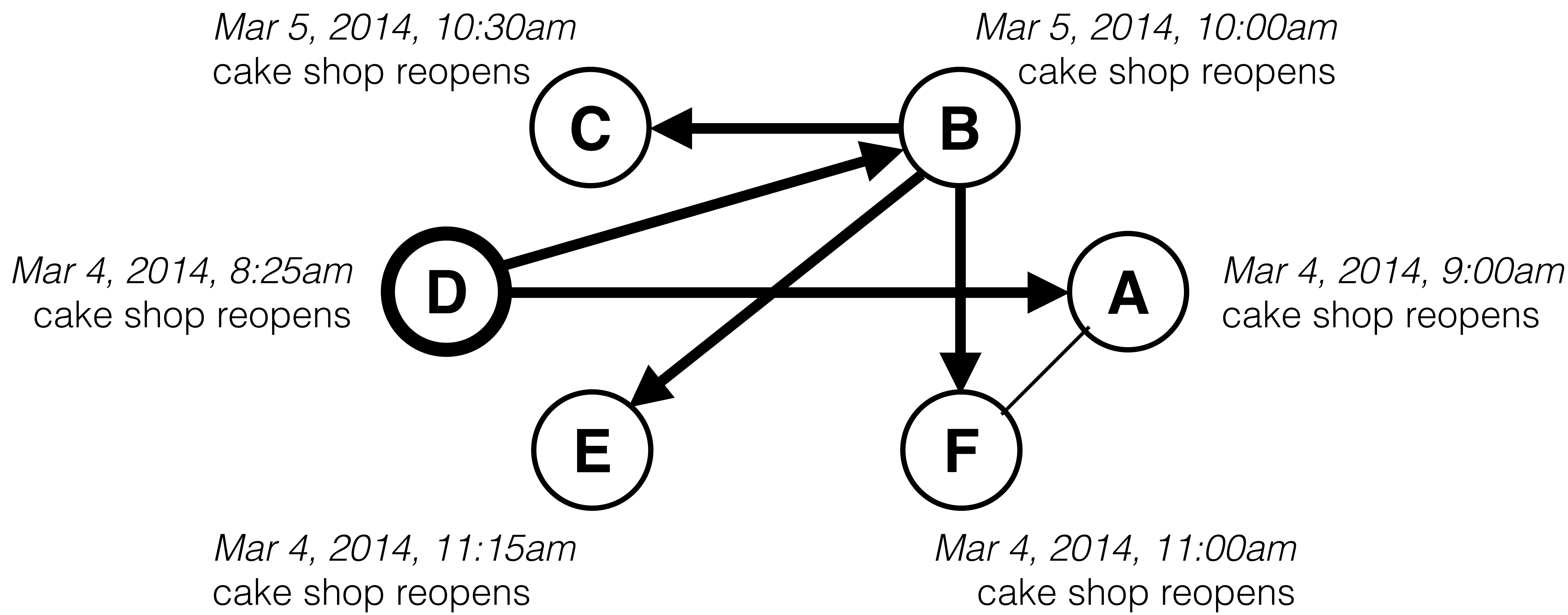


Mar 5, 2014, 10:30am
cake shop reopens
C
Mar 5, 2014, 10:00am
cake shop reopens
B
Mar 4, 2014, 8:25am
cake shop reopens
D
A
Mar 4, 2014, 9:00am
cake shop reopens
E
F
Mar 4, 2014, 11:00am
cake shop reopens

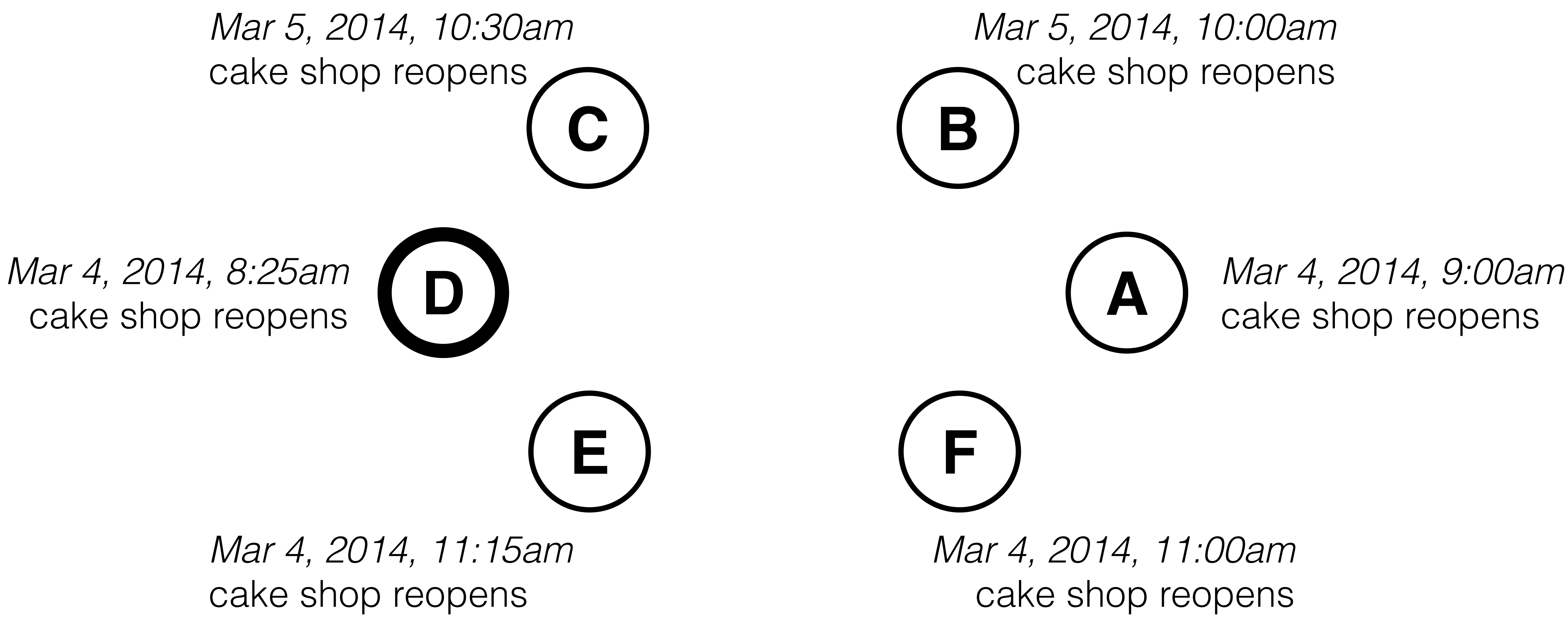
Meme Flow


Mar 5, 2014, 10:30am
cake shop reopens
Mar 5, 2014, 10:00am
cake shop reopens
C
B
Mar 4, 2014, 8:25am
cake shop reopens
D
A
Mar 4, 2014, 9:00am
cake shop reopens
E
F
Mar 4, 2014, 11:15am
cake shop reopens
Mar 4, 2014, 11:00am
cake shop reopens

Meme Flow

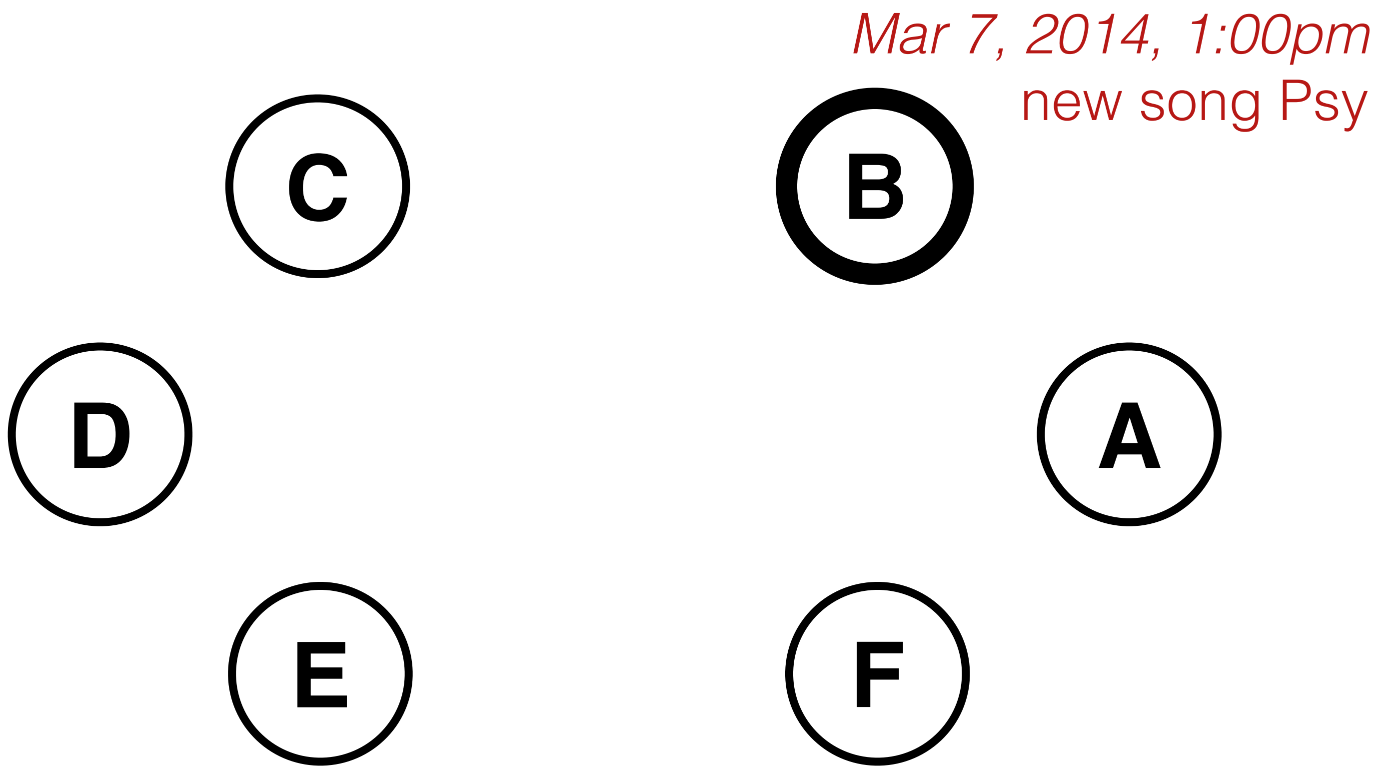

Mar 5, 2014, 10:30am
cake shop reopens
C
Mar 5, 2014, 10:00am
cake shop reopens
B
Mar 4, 2014, 8:25am
cake shop reopens
D
A
Mar 4, 2014, 9:00am
cake shop reopens
E
F
Mar 4, 2014, 11:15am
cake shop reopens
Mar 4, 2014, 11:00am
cake shop reopens

Traces



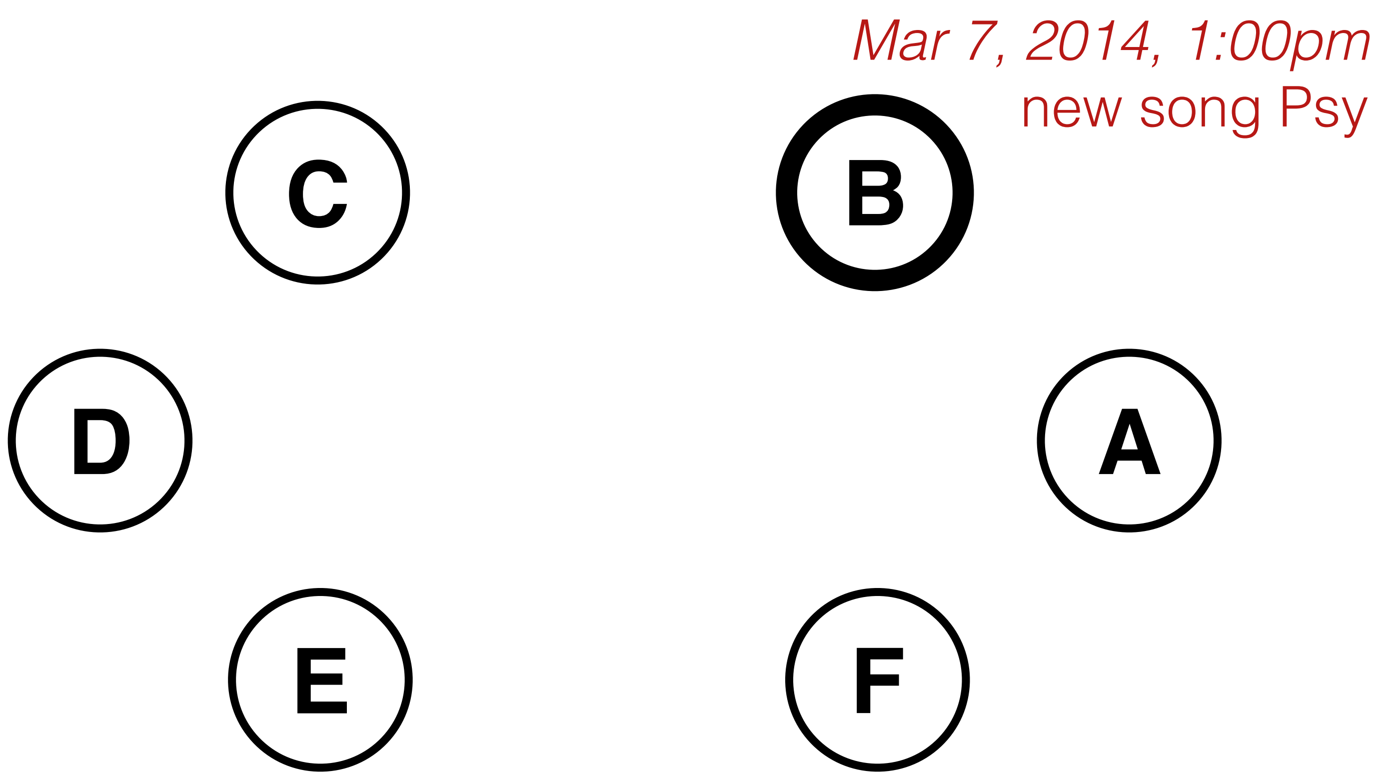
D *35’* A *60’* B *30’* C *30’* F *15’* E

Traces



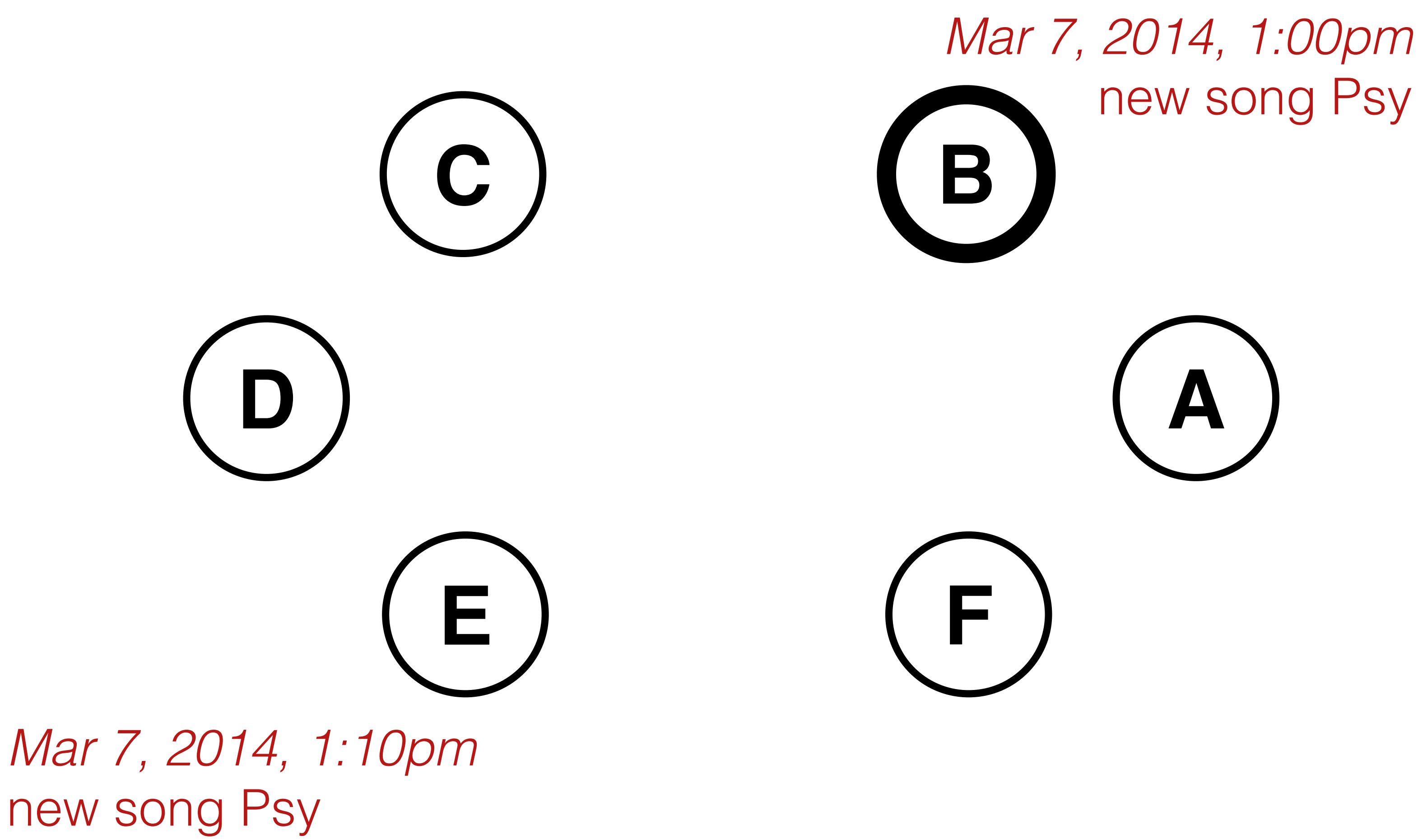
D *35'* A *60'* B *30'* C *30'* F *15'* E

Traces



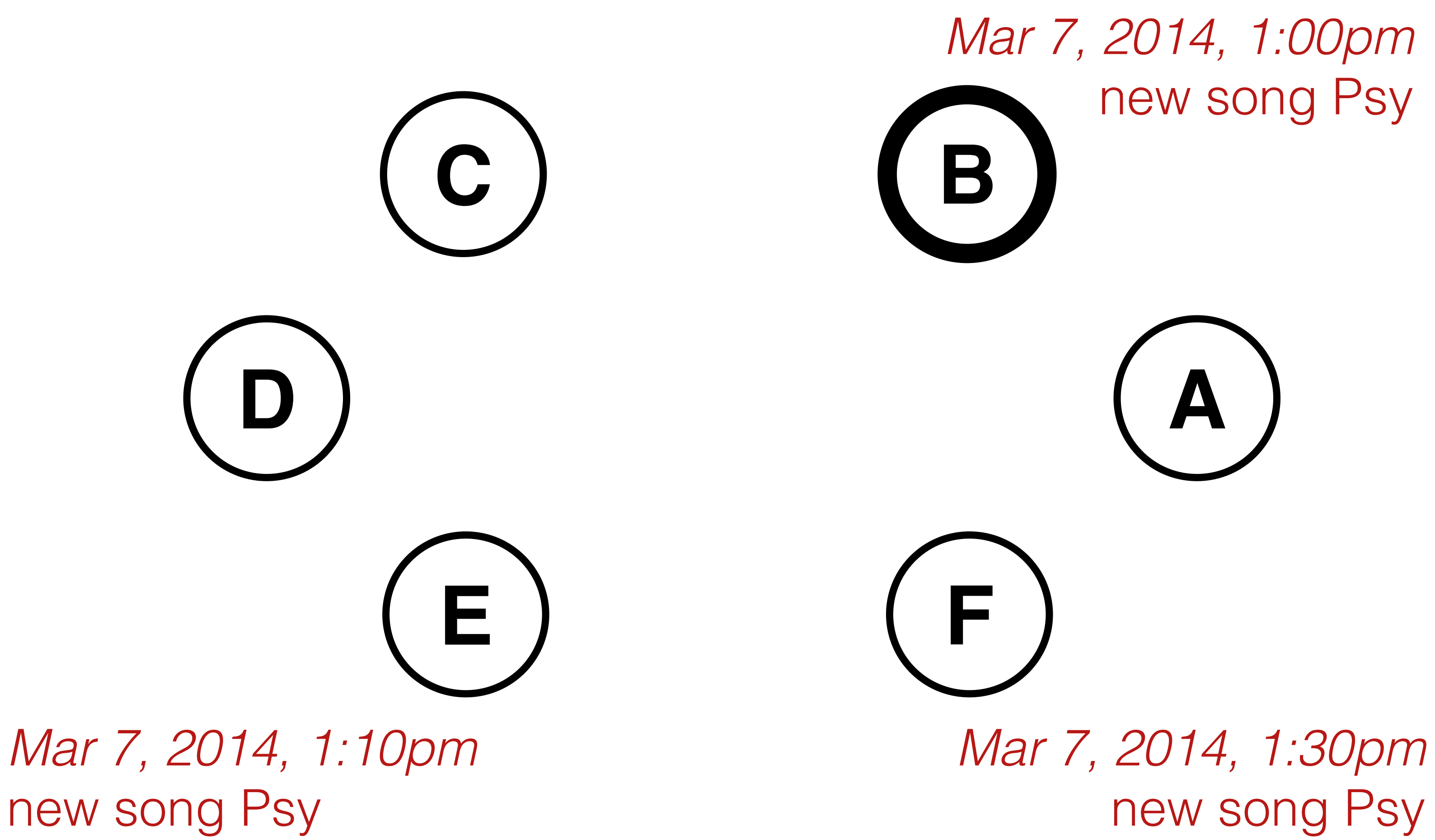
D *35'* A *60'* B *30'* C *30'* F *15'* E
B

Traces



D *35'* A *60'* B *30'* C *30'* F *15'* E
B *10'* E

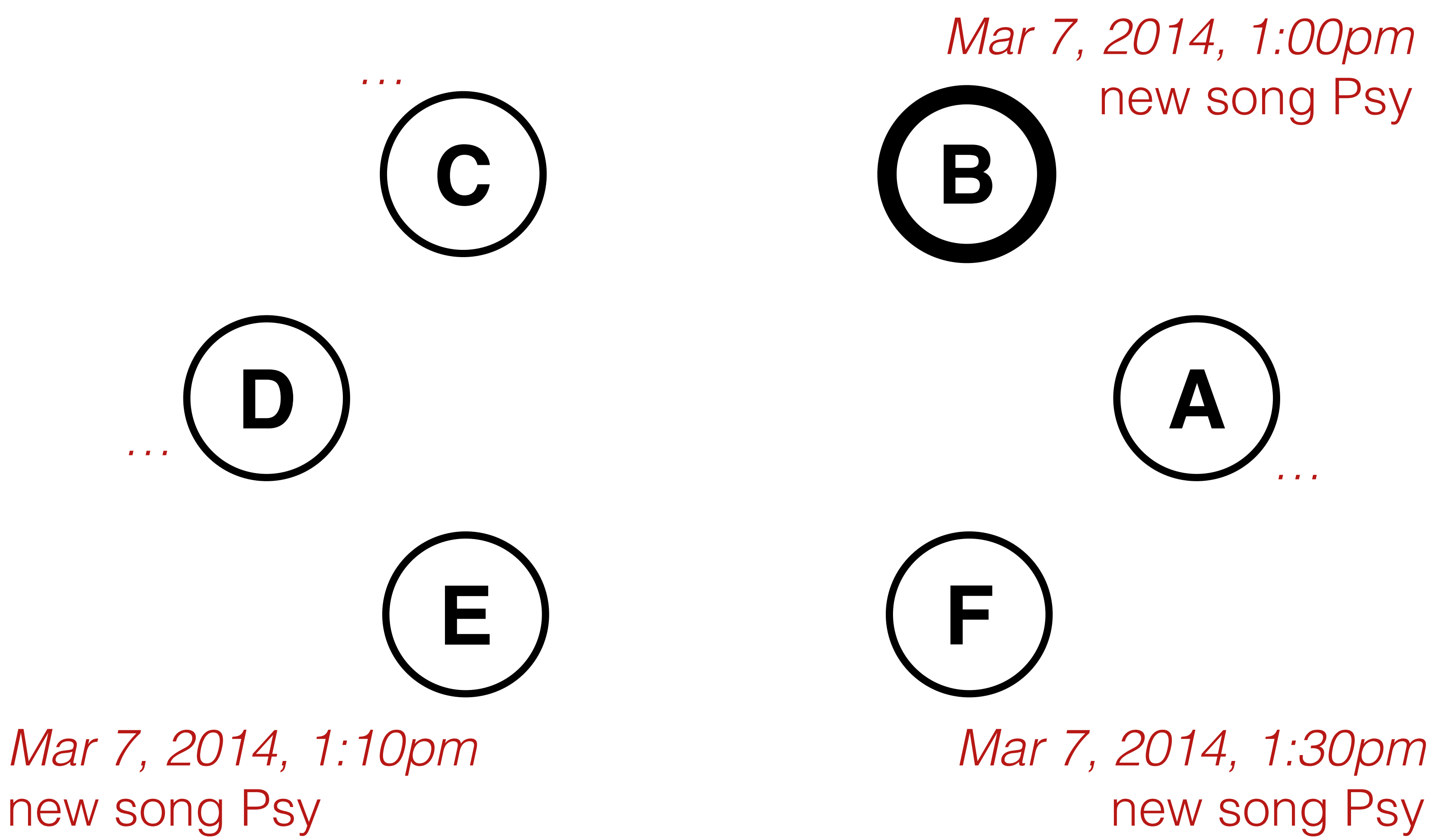
Traces



D *35’* A *60’* B *30’* C *30’* F *15’* E

B *10’* E *20’* F

Traces



D *35'* A *60'* B *30'* C *30'* F *15'* E

B *10'* E *20'* F *15'* A *15'* D *30'* C

Inference

Given a set of traces

D	*35'*	A	*60'*	B	*30'*	C	*30'*	F	*15'*	E
B	*10'*	E	*20'*	F	*15'*	A	*15'*	D	*30'*	C
B	*12'*	D	*15'*	A	*10'*	C	*25'*	E	*30'*	F

...

Inference

Given a set of traces

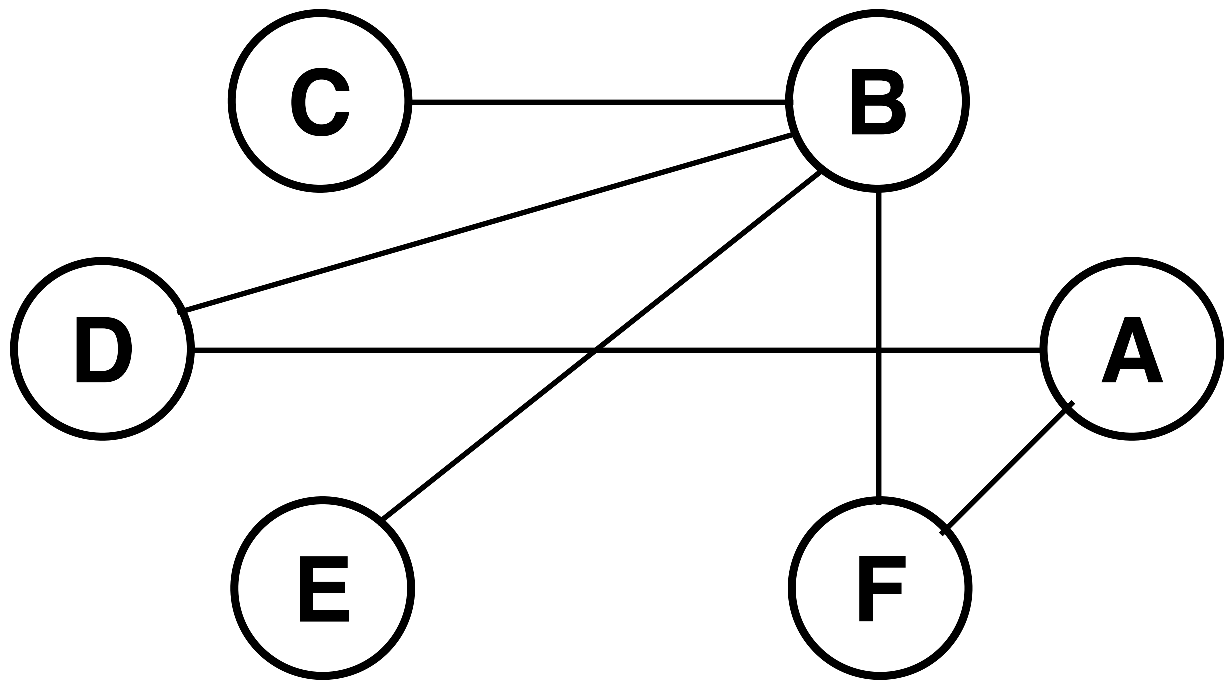
D	*35'*	A	*60'*	B	*30'*	C	*30'*	F	*15'*	E
B	*10'*	E	*20'*	F	*15'*	A	*15'*	D	*30'*	C
B	*12'*	D	*15'*	A	*10'*	C	*25'*	E	*30'*	F

...

we would like to infer the
unknown blogger graph *G*

Trace Spreading Model

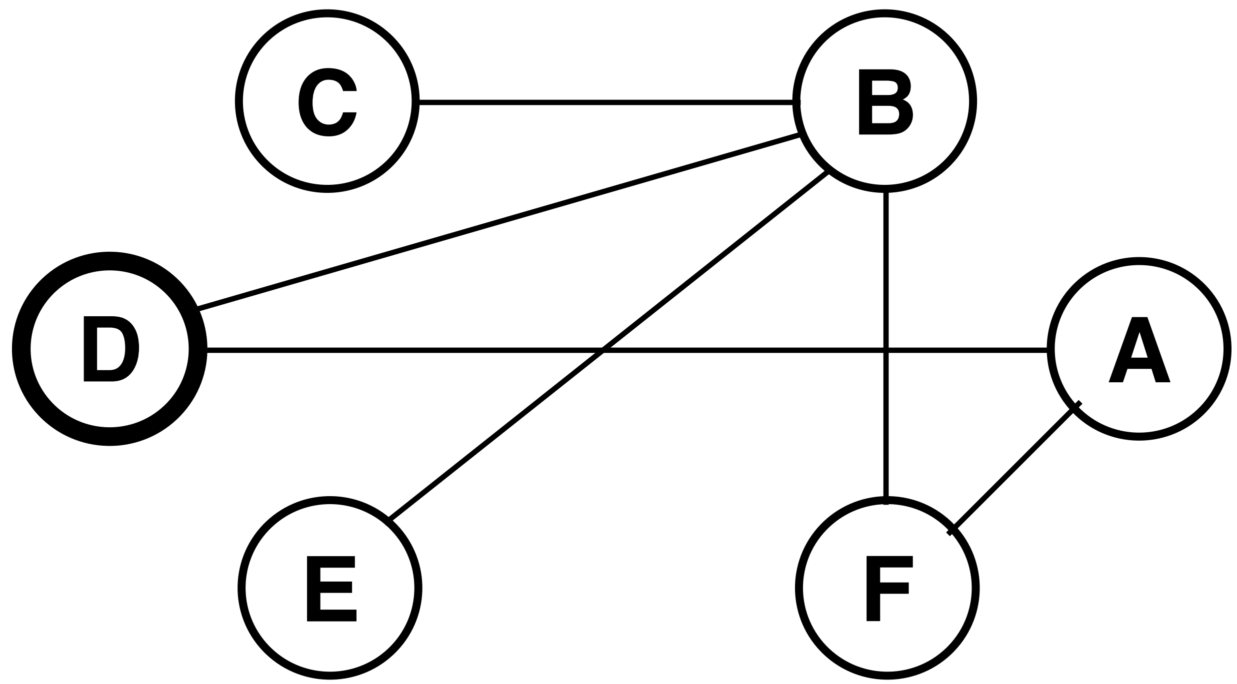
Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



- The unique source is chosen uniformly at random

Trace Spreading Model

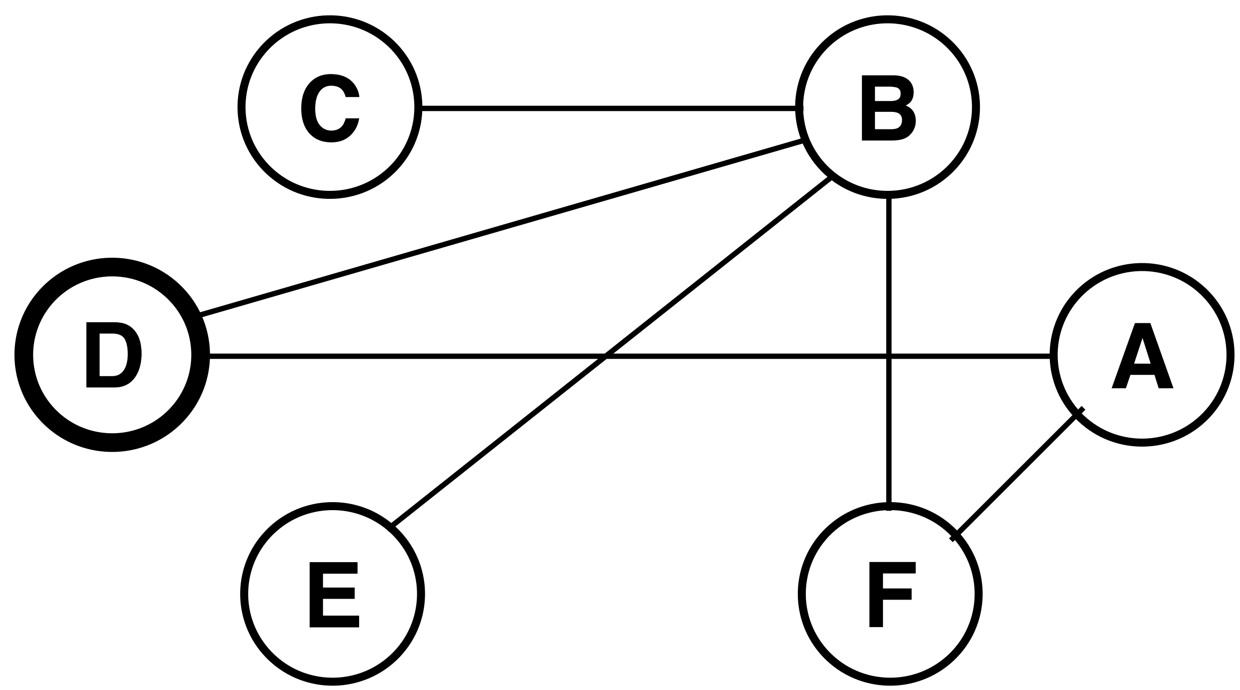
Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



- The unique source is chosen uniformly at random

Trace Spreading Model

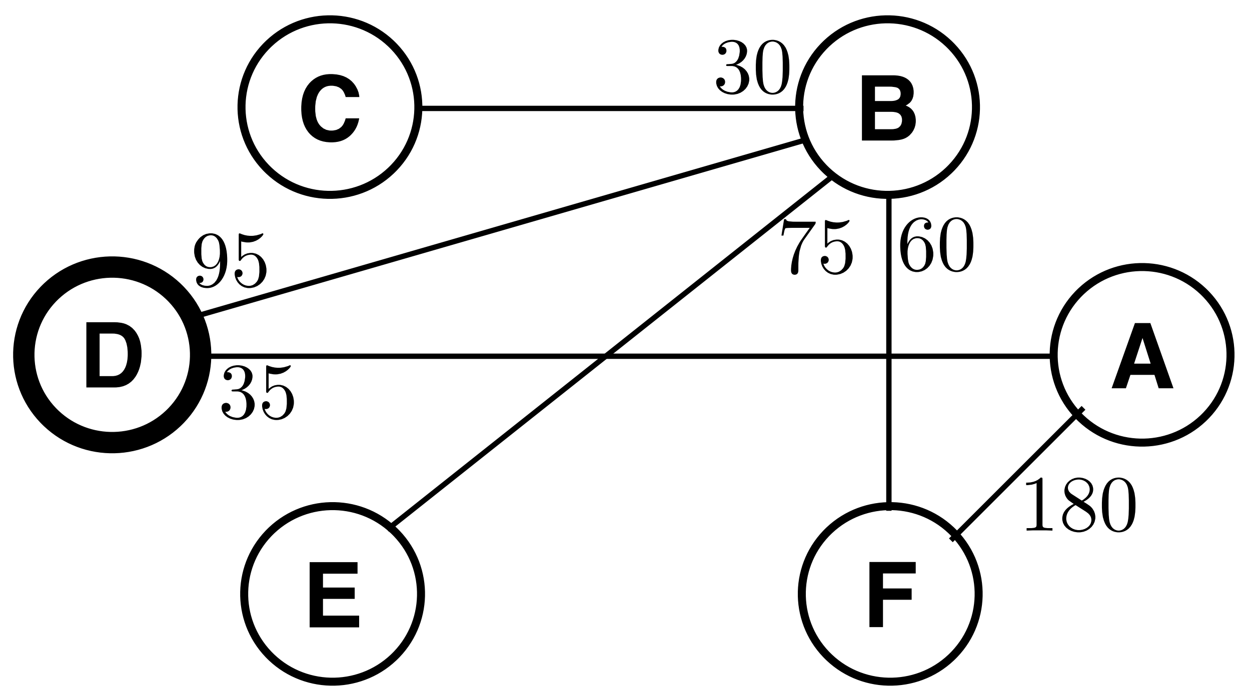
Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



- The unique source is chosen uniformly at random
- The time to traverse an edge is an iid sample of $\mathrm{Exp}(\lambda)$

Trace Spreading Model

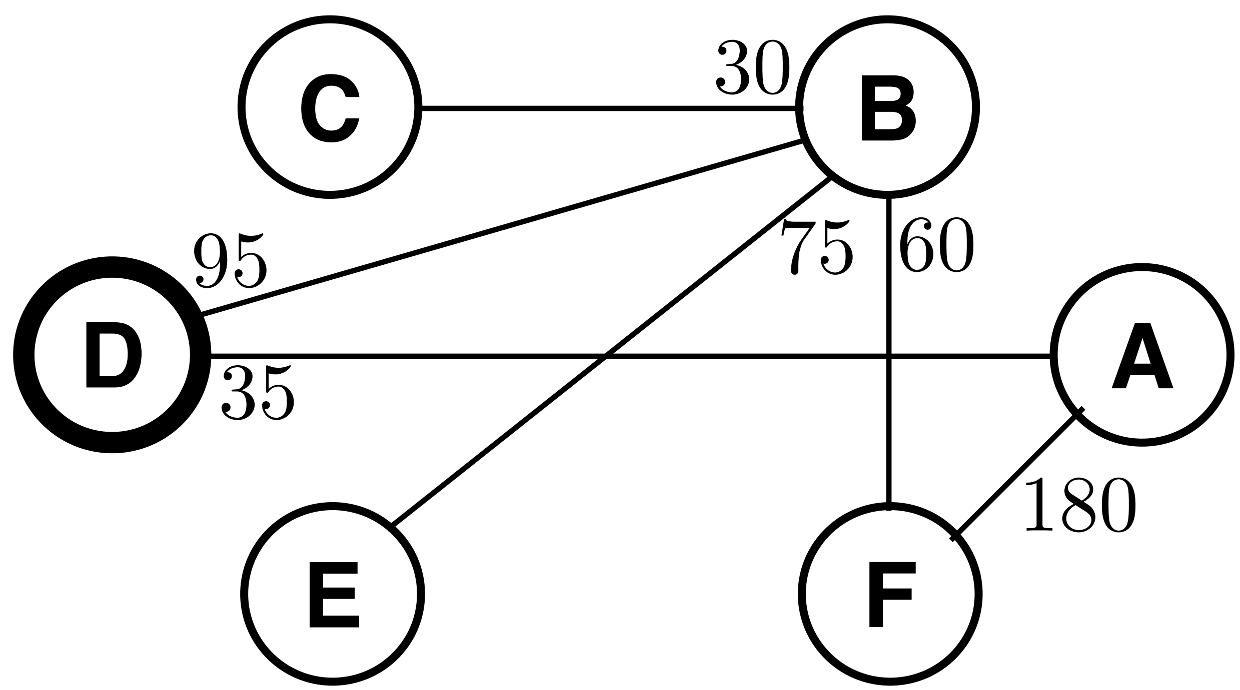
Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



- The unique source is chosen uniformly at random
- The time to traverse an edge is an iid sample of $\mathrm{Exp}(\lambda)$

Trace Spreading Model

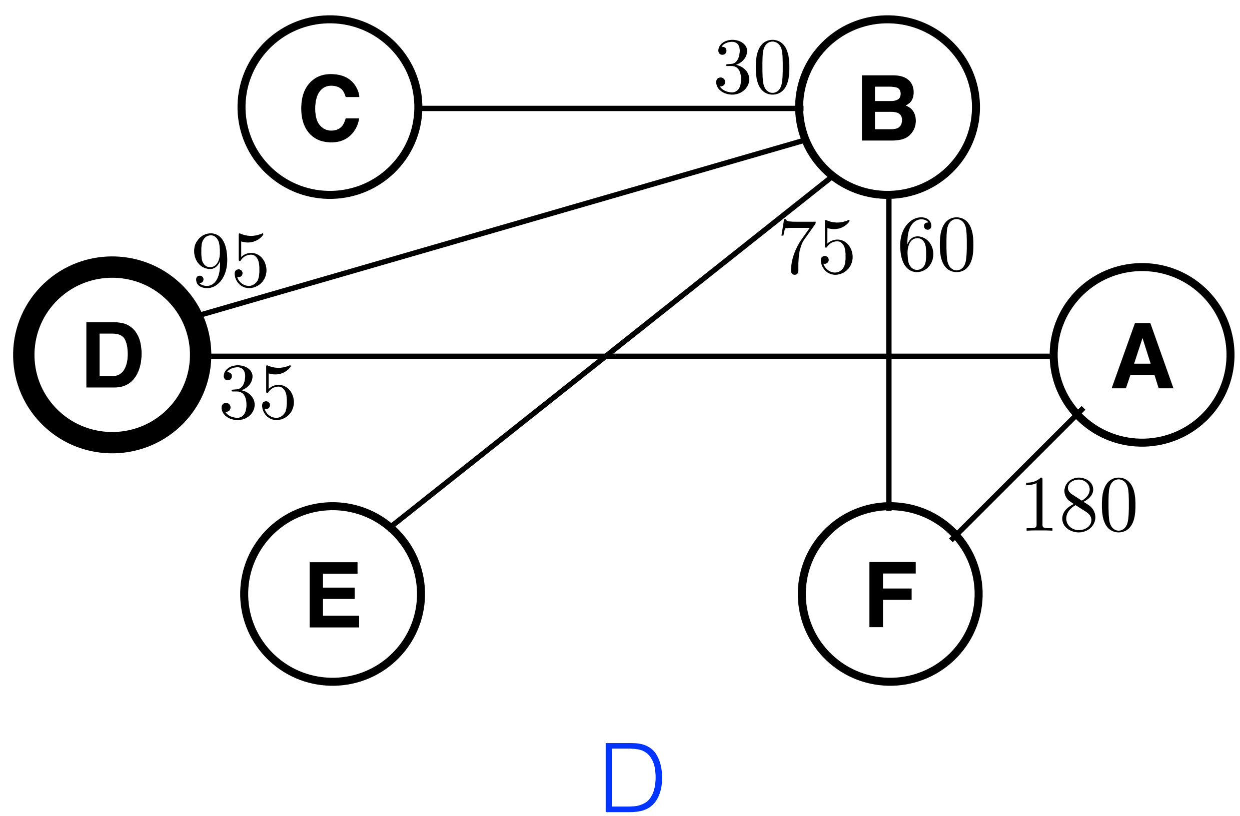
Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



- The unique source is chosen uniformly at random
- The time to traverse an edge is an iid sample of $\mathrm{Exp}(\lambda)$
- The trace follows the shortest path tree from the source

Trace Spreading Model

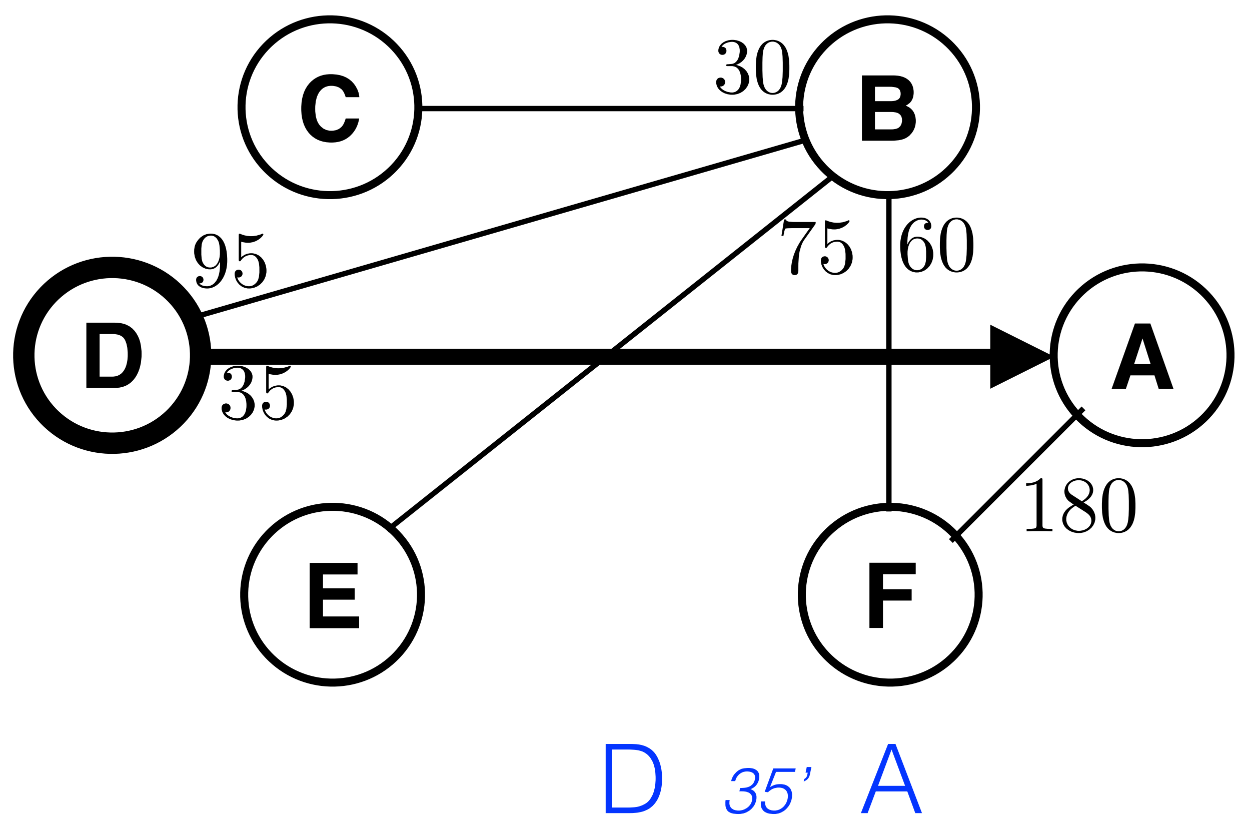
Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



- The unique source is chosen uniformly at random
- The time to traverse an edge is an iid sample of $\mathrm{Exp}(\lambda)$
- The trace follows the shortest path tree from the source

Trace Spreading Model

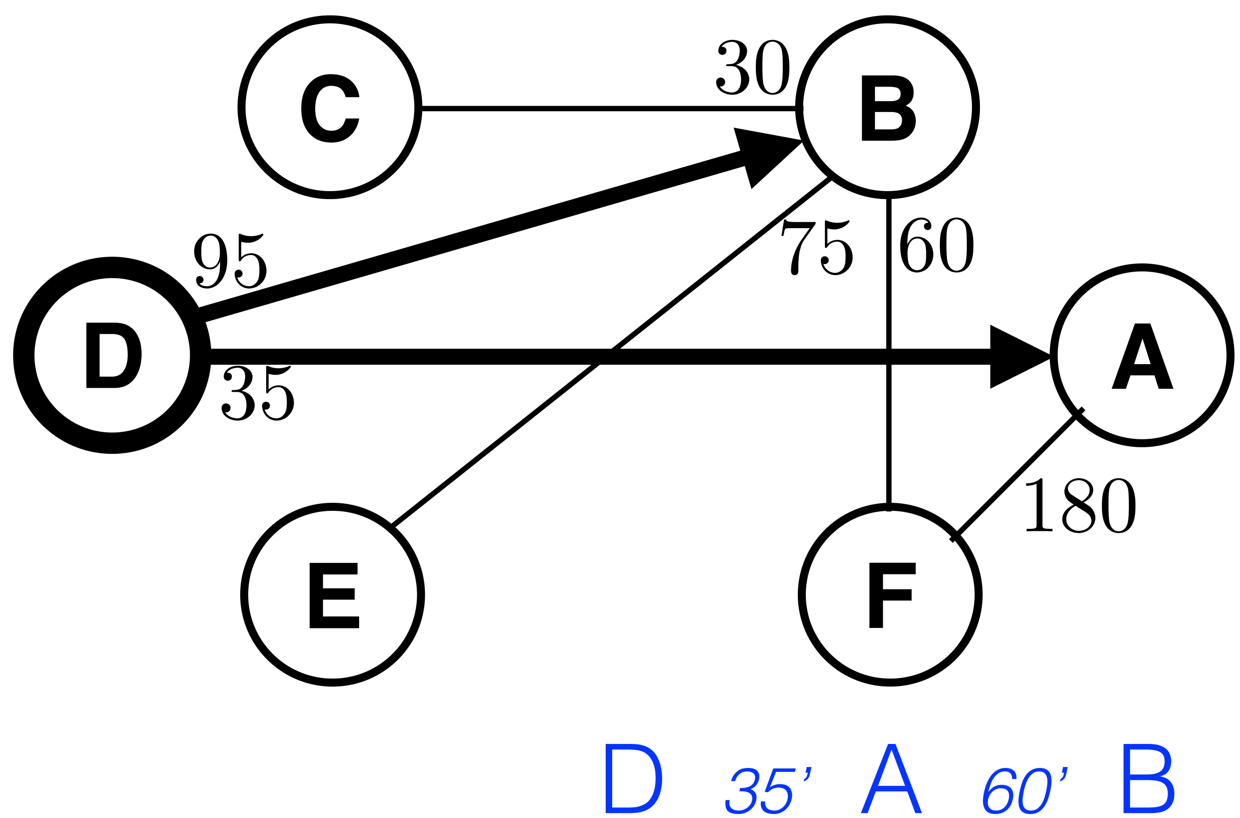
Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



- The unique source is chosen uniformly at random
- The time to traverse an edge is an iid sample of $\mathrm{Exp}(\lambda)$
- The trace follows the shortest path tree from the source

Trace Spreading Model

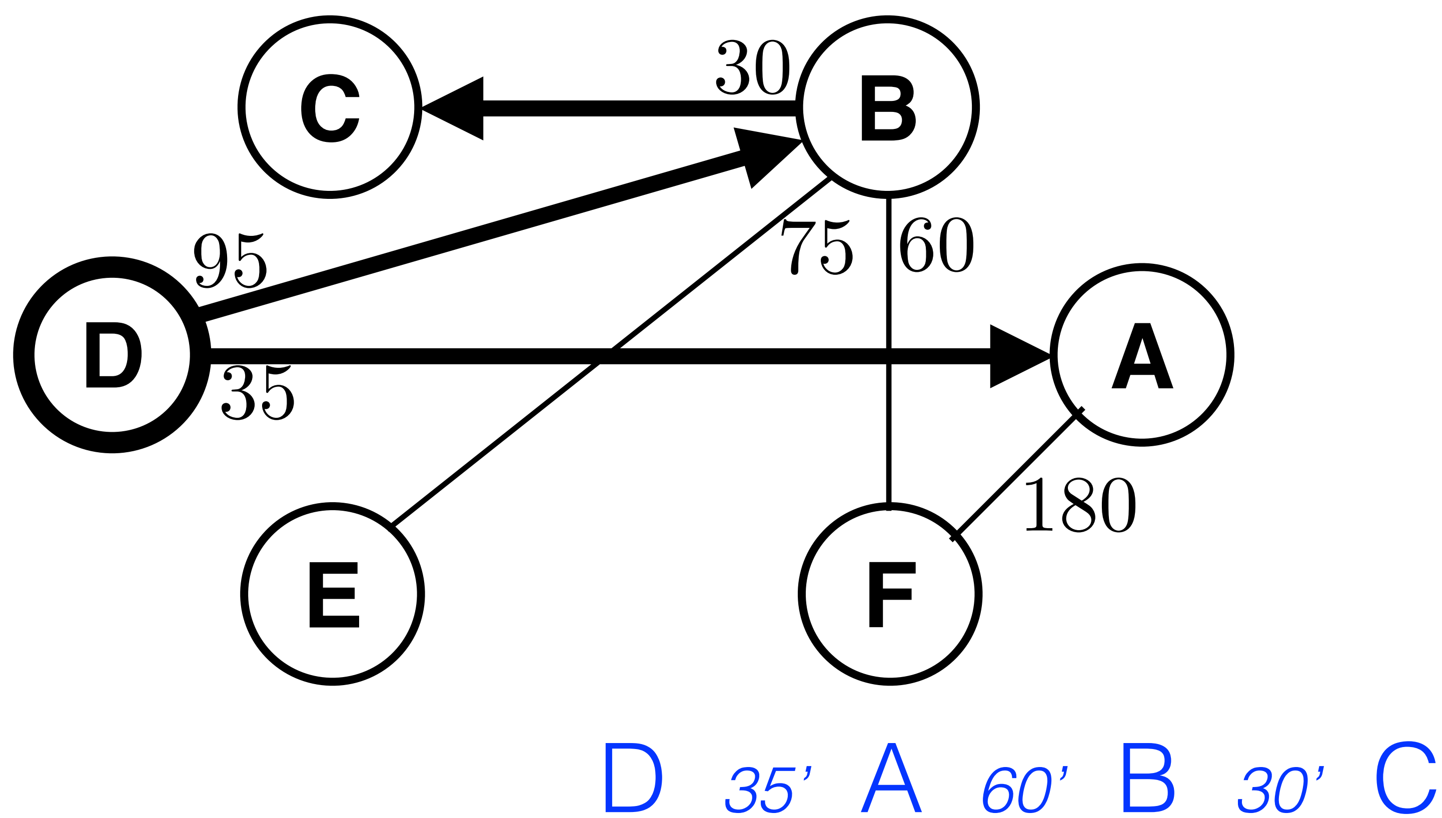
Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



- The unique source is chosen uniformly at random
- The time to traverse an edge is an iid sample of $\mathrm{Exp}(\lambda)$
- The trace follows the shortest path tree from the source

Trace Spreading Model

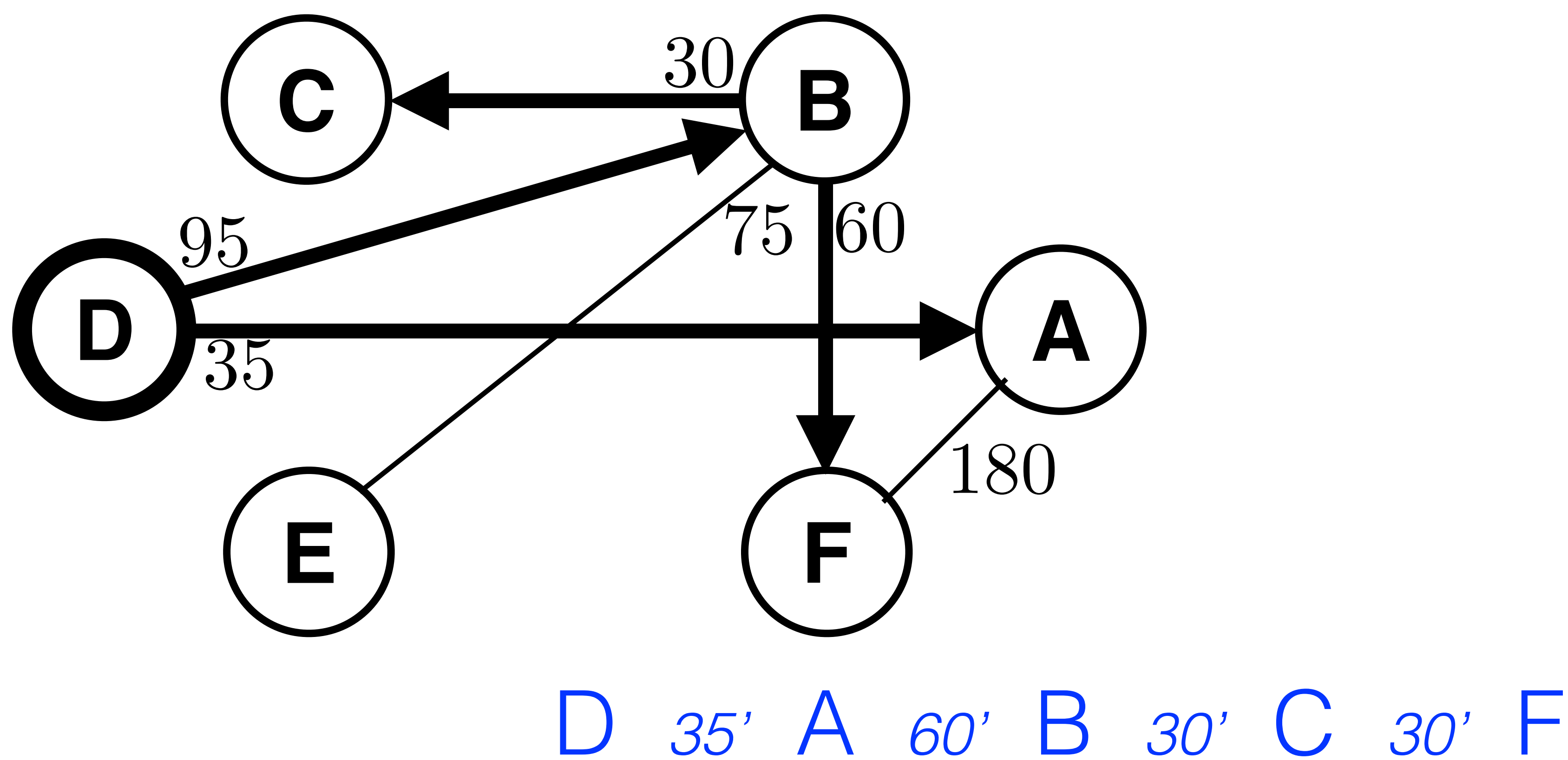
Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



- The unique source is chosen uniformly at random
- The time to traverse an edge is an iid sample of $\mathrm{Exp}(\lambda)$
- The trace follows the shortest path tree from the source

Trace Spreading Model

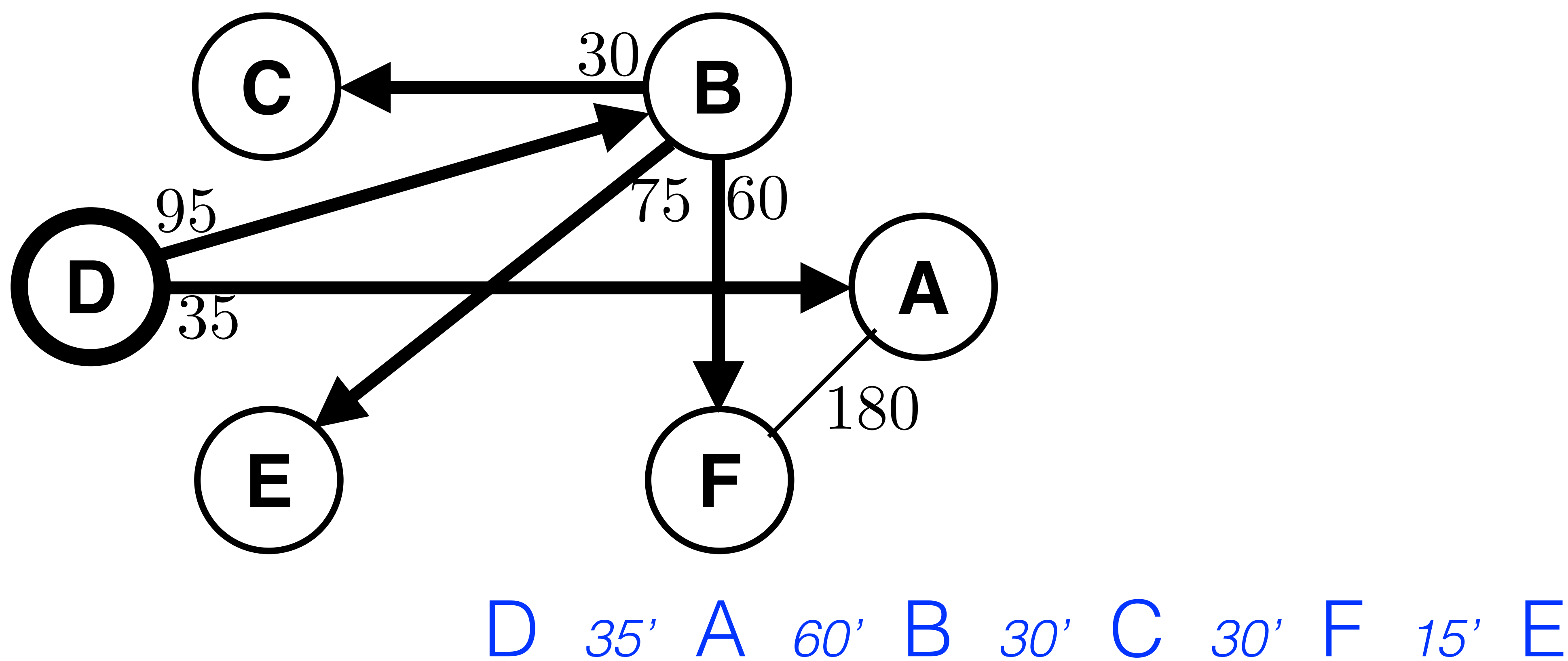
Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



- The unique source is chosen uniformly at random
- The time to traverse an edge is an iid sample of $\mathrm{Exp}(\lambda)$
- The trace follows the shortest path tree from the source

Trace Spreading Model

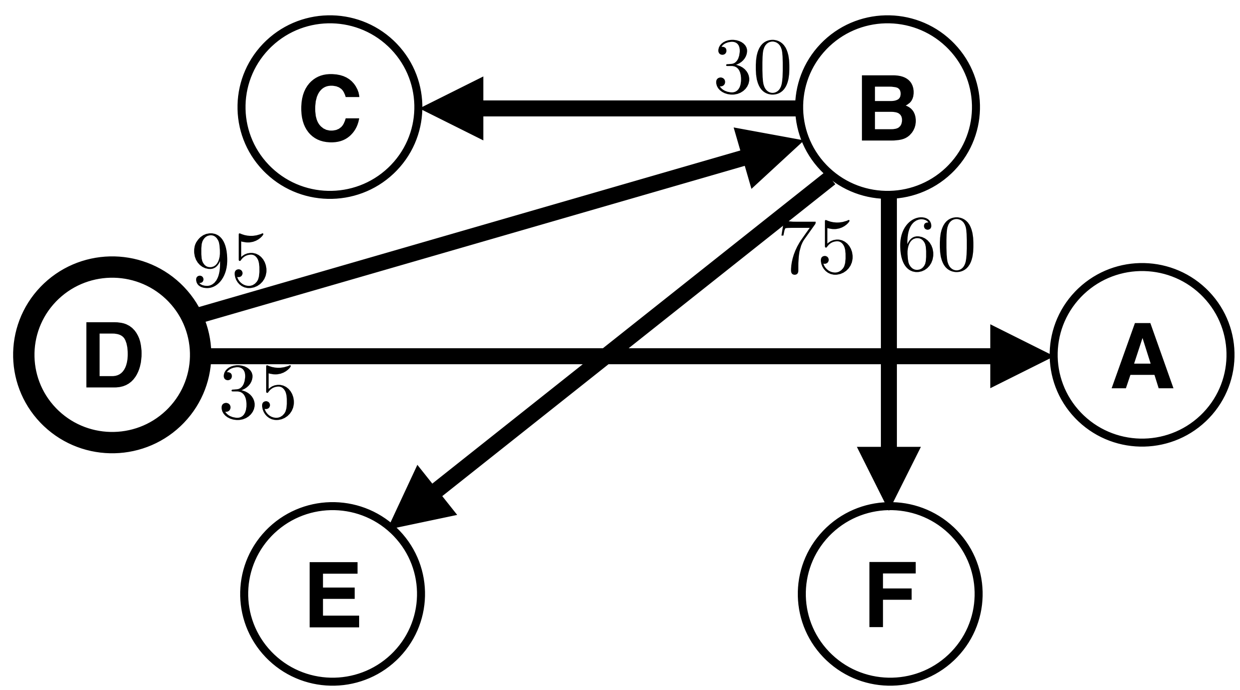
Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



- The unique source is chosen uniformly at random
- The time to traverse an edge is an iid sample of $\mathrm{Exp}(\lambda)$
- The trace follows the shortest path tree from the source

Trace Spreading Model

Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*

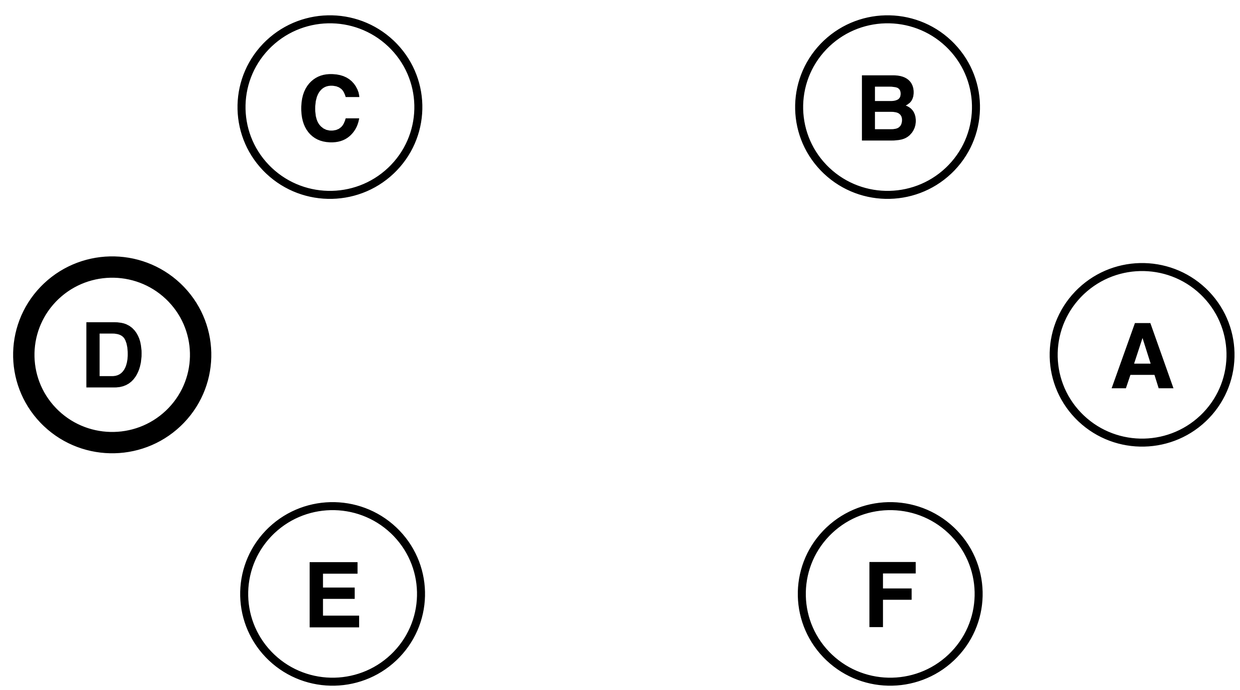


D *35'* A *60'* B *30'* C *30'* F *15'* E

- The unique source is chosen uniformly at random
- The time to traverse an edge is an iid sample of $\mathrm{Exp}(\lambda)$
- The trace follows the shortest path tree from the source

Trace Spreading Model

Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*



D *35'* A *60'* B *30'* C *30'* F *15'* E

- The unique source is chosen uniformly at random
- The time to traverse an edge is an iid sample of $\mathrm{Exp}(\lambda)$
- The trace follows the shortest path tree from the source

Can we make a sensible guess of the network?

Can we make a sensible guess of the network?

How many traces would we need?

Our Contributions

Abrahao, Chierichetti, Kleinberg, Panconesi *[KDD'13]*

- We show that $O\left(n\Delta \log n\right)$ traces are sufficient for reconstruction.
- We show that, in some cases, $\Omega\left(\frac{n\Delta}{\log^2 n}\right)$ traces are necessary.

n is the number of nodes of the unknown graph,
Δ is its maximum degree.

Our Contributions

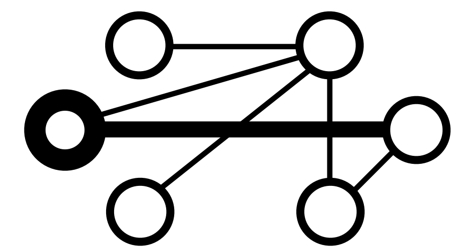
Abrahao, Chierichetti, Kleinberg, Panconesi *[KDD'13]*

- We show that $O\left(n\Delta\log n\right)$ traces are sufficient for reconstruction.
- We show that, in some cases, $\Omega\left(\frac{n\Delta}{\log^2 n}\right)$ traces are necessary.
- We also show that $O\left(\mathrm{poly}(\Delta)\cdot\log n\right)$ traces are sufficient, that $O(\log n)$ traces are sufficient for trees…

A simple algorithm

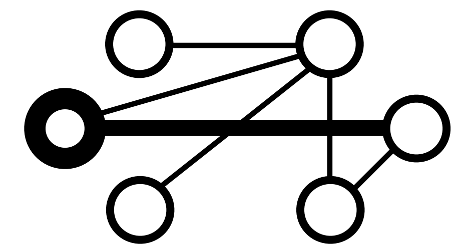
- **Observation**
  If u and v are the first two nodes of a trace then the edge $\{u,v\}$ is in the unknown graph.

A simple algorithm



- **Observation**
  If *u* and *v* are the first two nodes of a trace then the edge *{u,v}* is in the unknown graph.

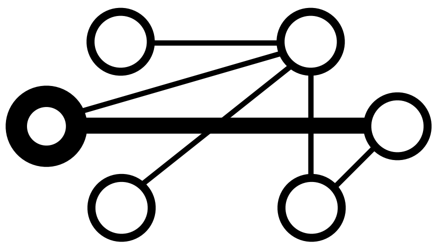
A simple algorithm



- **Observation**
  If *u* and *v* are the first two nodes of a trace then the edge *{u,v}* is in the unknown graph.

- We can then perfectly reconstruct if, for each unknown edge *{u,v}*, there exists a trace that begins with its two endpoints.

A simple algorithm

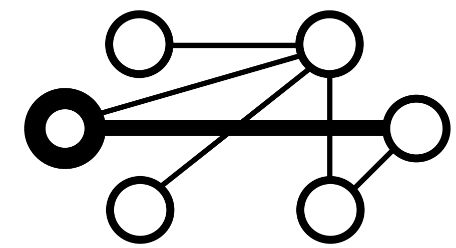


- **Observation**
  If *u* and *v* are the first two nodes of a trace then the edge *{u,v}* is in the unknown graph.

- We can then perfectly reconstruct if, for each unknown edge *{u,v}*, there exists a trace that begins with its two endpoints.

D *35'* A *60'* B *30'* C *30'* F *15'* E
B *10'* E *20'* F *15'* A *15'* D *30'* C
B *12'* D *15'* A *10'* C *25'* E *30'* F

A simple algorithm



- **Observation**
  If *u* and *v* are the first two nodes of a trace then the edge *{u,v}* is in the unknown graph.

- We can then perfectly reconstruct if, for each unknown edge *{u,v}*, there exists a trace that begins with its two endpoints.

D *35'* A *60'* B *30'* C *30'* F *15'* E
B *10'* E *20'* F *15'* A *15'* D *30'* C
B *12'* D *15'* A *10'* C *25'* E *30'* F

A simple algorithm

- The probability that a specific node is the first one in a trace is ${}^{1}\!/_{n}$.

- The probability that the edge {u,v} is spanned by the first two nodes of the trace is then at least

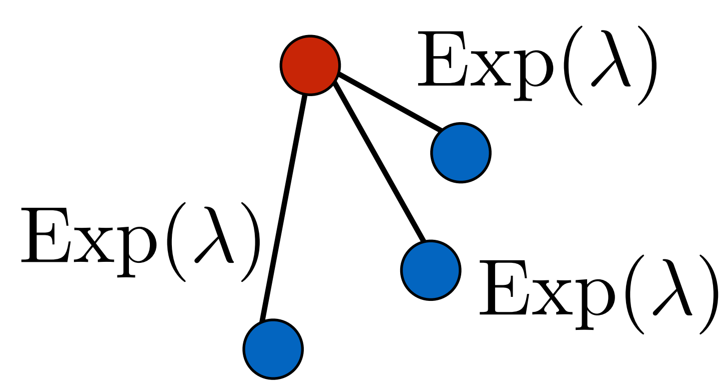
$$\frac{1}{n \cdot \min(\deg(u), \deg(v))}$$

A simple algorithm

- The probability that a specific node is the first one in a trace is ${}^1\!/_n$.

- The probability that the edge {u,v} is spanned by the first two nodes of the trace is then at least

$$\frac{1}{n \cdot \min(\deg(u), \deg(v))}$$



A simple algorithm

- The probability that a specific node is the first one in a trace is $1/n$.

- The probability that the edge {u,v} is spanned by the first two nodes of the trace is then at least
$$\frac{1}{n \cdot \min(\deg(u), \deg(v))}$$

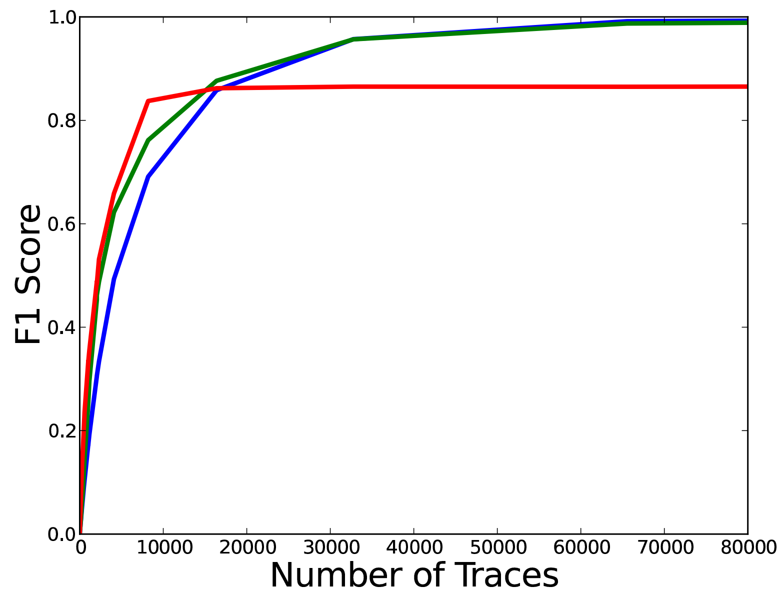
- Any classical tail bound then proves that $O(n\Delta \log n)$ traces are enough to perfectly reconstruct the graph.

The first-edge algorithm

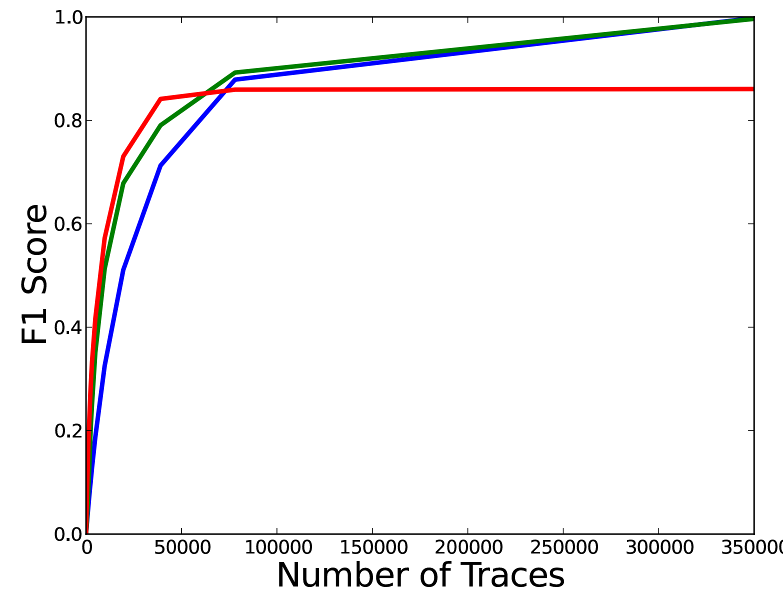
- Works for any (uniform) edge waiting time distribution.

- It is competitive with more complex machine-learning based algorithms (Gomez-Rodriguez, Leskovec, Krause *[KDD'10]*).

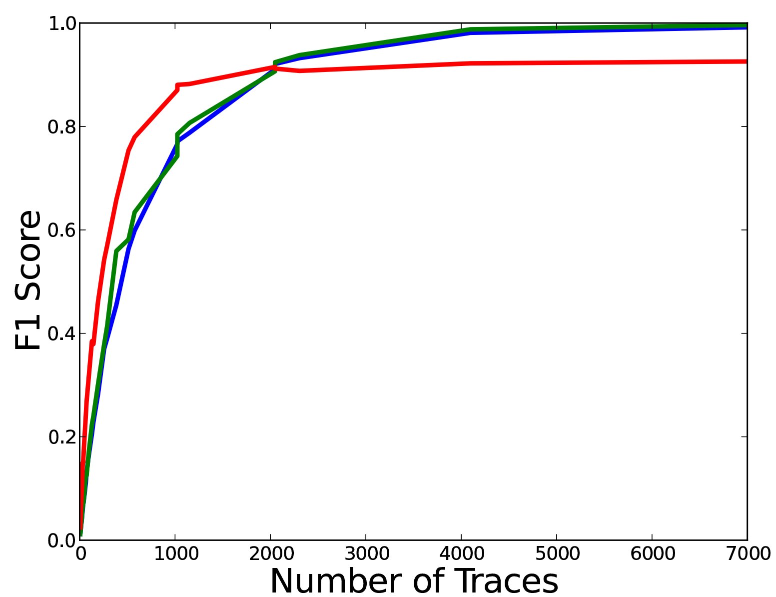
Performances



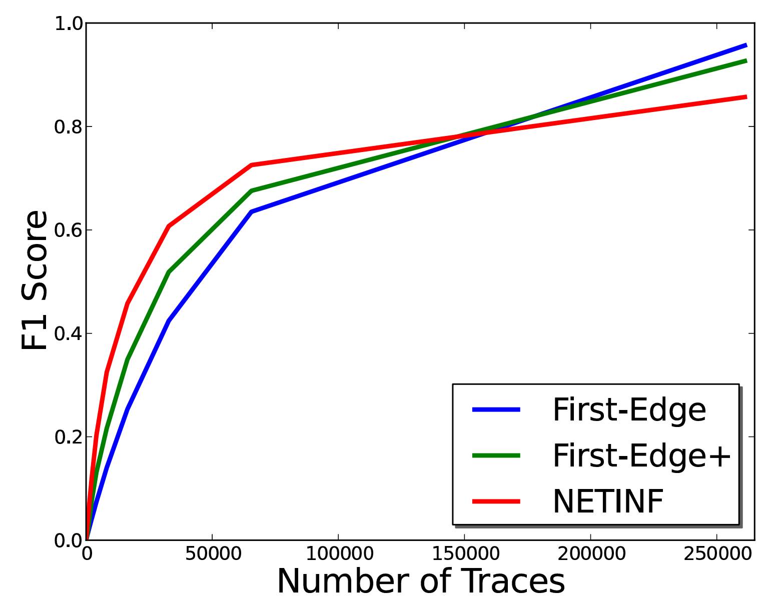
(a) Barabasi-Albert



(b) Facebook-Rice Undergrad



(c) Power-Law Tree



(d) $G_{n,p}$

Our Contributions

Abrahao, Chierichetti, Kleinberg, Panconesi *[KDD'13]*

- We show that $O\left(n\Delta\log n\right)$ traces are sufficient for reconstruction.
- We show that, in some cases, $\Omega\left(\frac{n\Delta}{\log^2 n}\right)$ traces are necessary.

Our Contributions

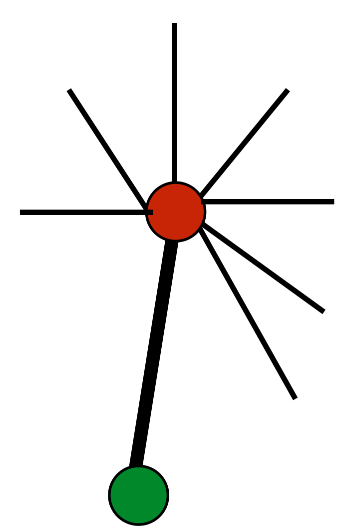
Abrahao, Chierichetti, Kleinberg, Panconesi *[KDD'13]*

- We show that $O(nD \log n)$ traces are sufficient for guessing the edges incident on a node of degree $\leq D$
- We show that, in some cases, $\Omega\left(\frac{n\Delta}{\log^2 n}\right)$ traces are necessary.

Our Contributions

Abrahao, Chierichetti, Kleinberg, Panconesi *[KDD'13]*

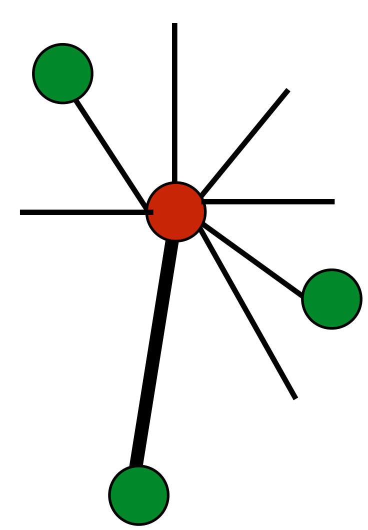
- We show that $O(nD \log n)$ traces are sufficient for guessing the edges incident on a node of degree $\leq D$
- We show that, in some cases, $\Omega\left(\frac{n\Delta}{\log^2 n}\right)$ traces are necessary.



Our Contributions

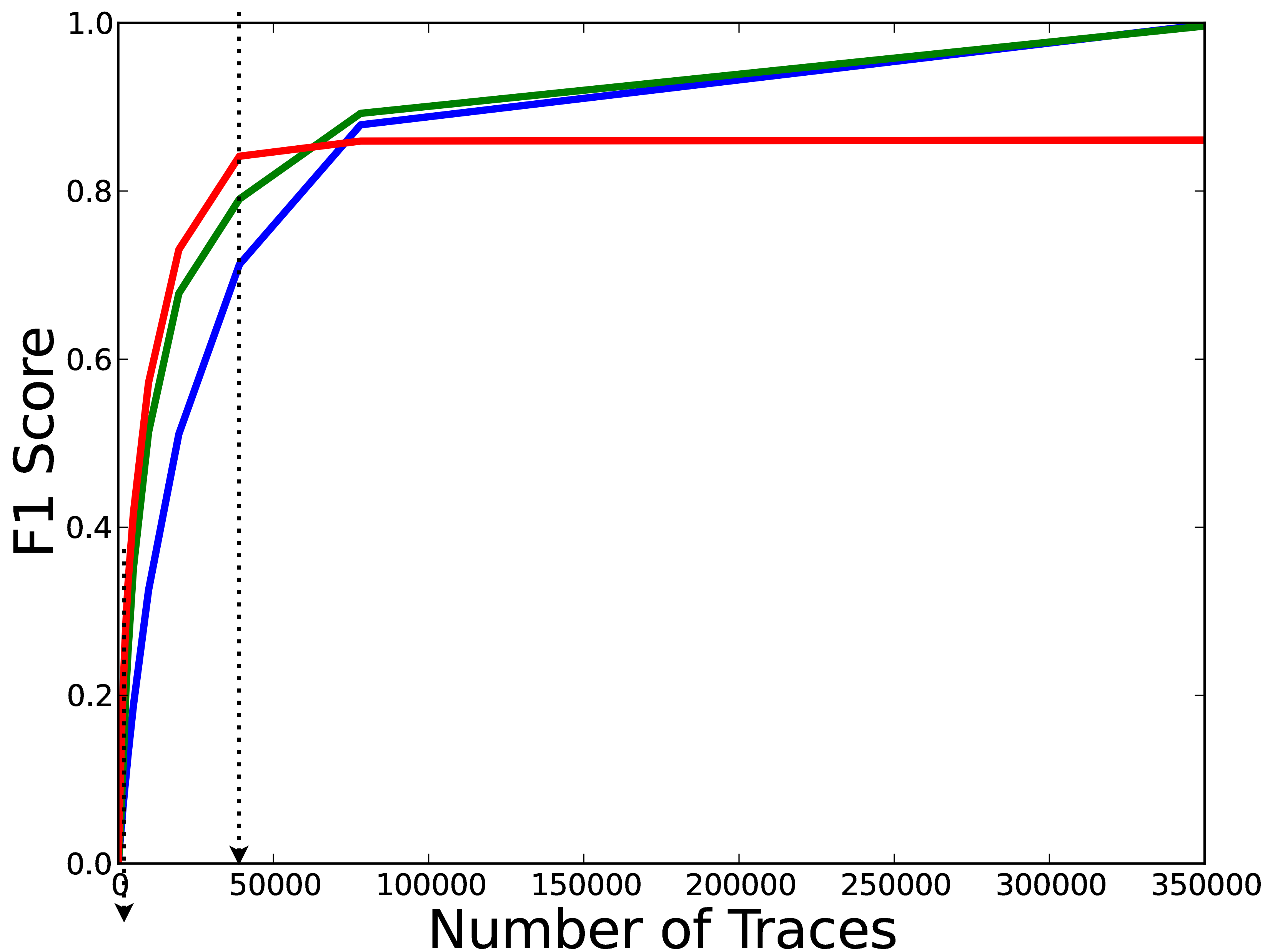
Abrahao, Chierichetti, Kleinberg, Panconesi *[KDD'13]*

- We show that $O(nD \log n)$ traces are sufficient for guessing the edges incident on a node of degree $\leq D$
- We show that, in some cases, $\Omega\left(\frac{n\Delta}{\log^2 n}\right)$ traces are necessary.



Why ~ 40K traces?

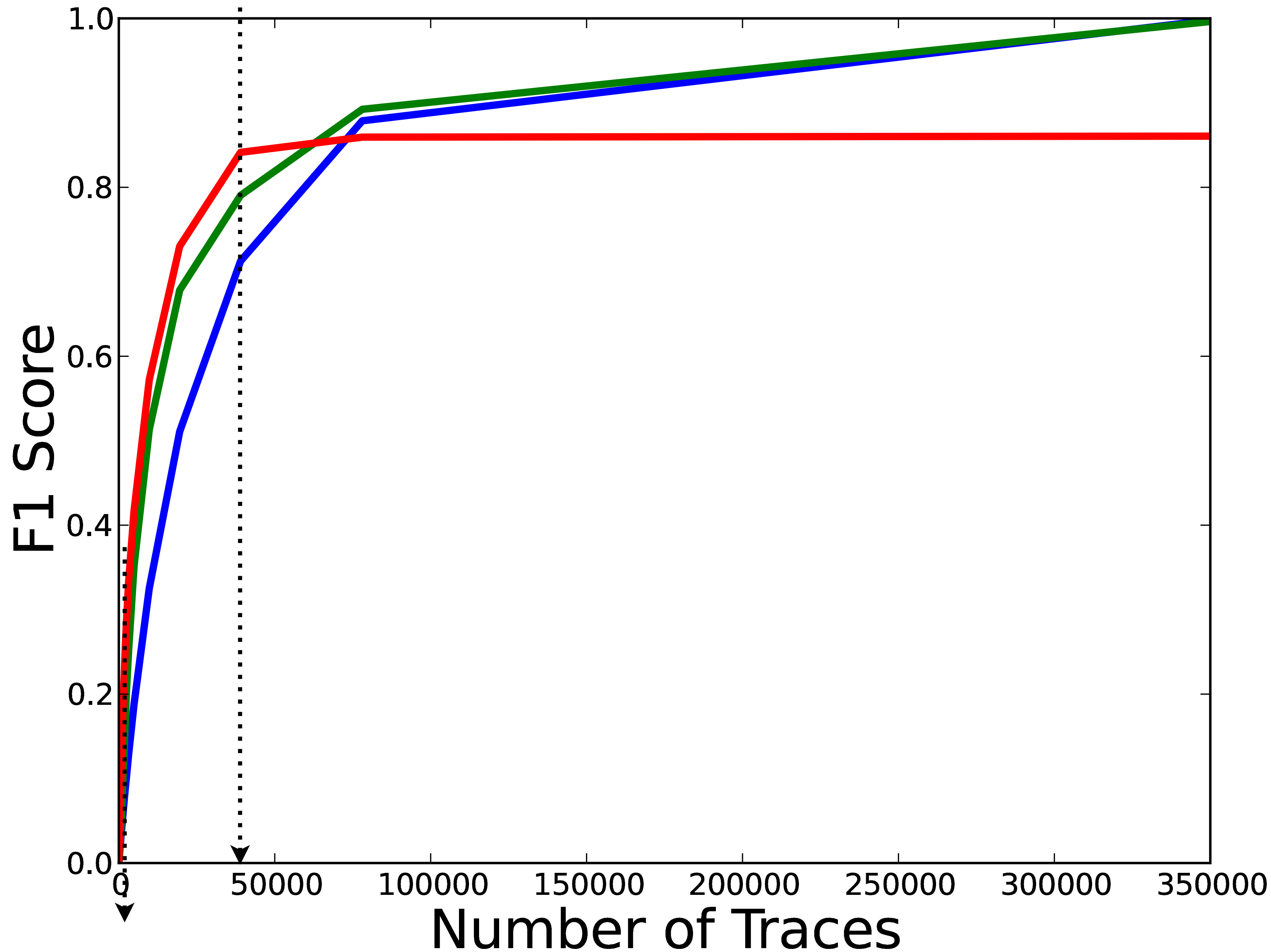
Most edges are incident on nodes having small degree



Facebook-Rice Undergrad

Why ~ 40K traces?

Most edges are incident on nodes having small degree

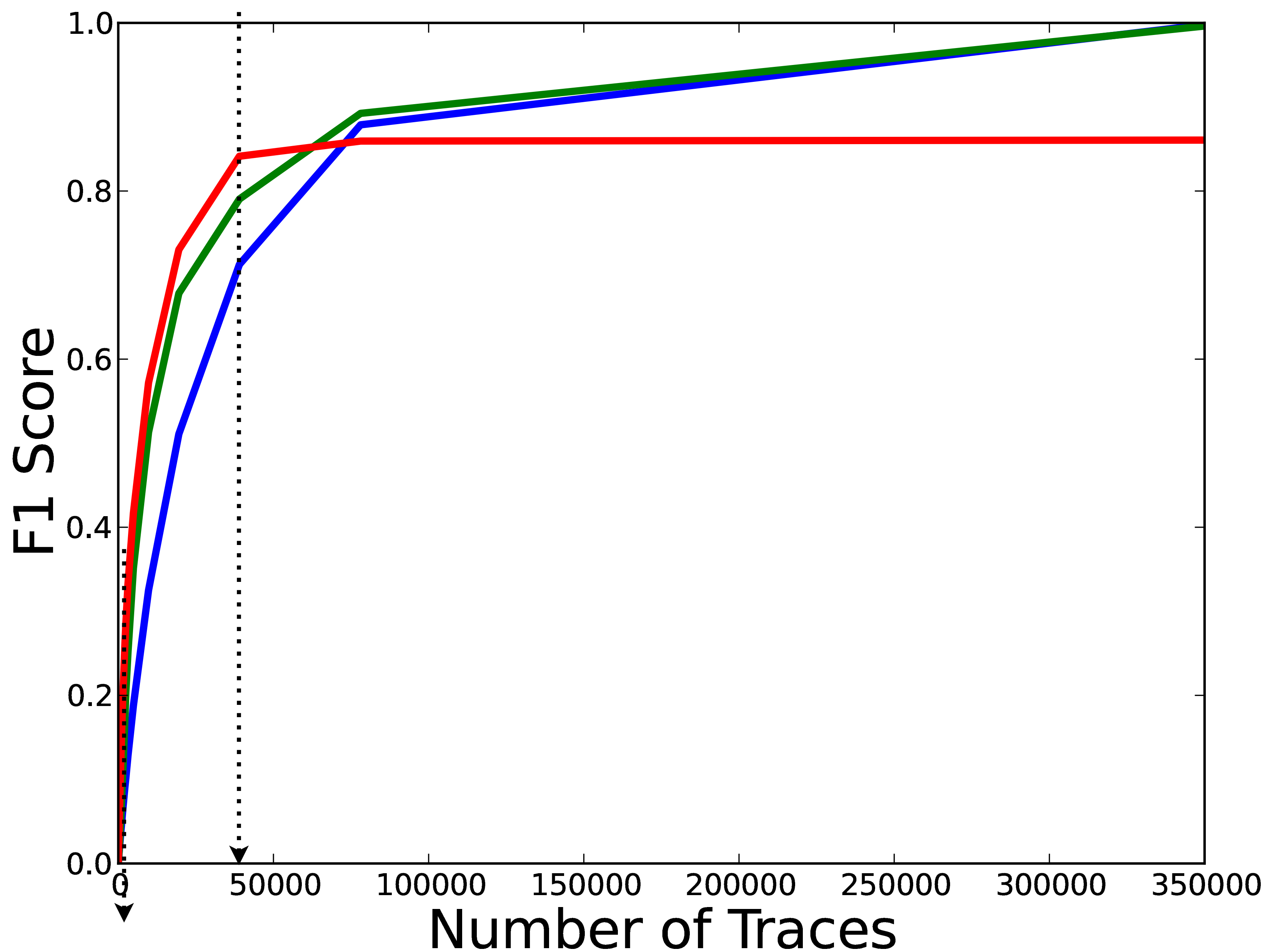


Facebook-Rice Undergrad

Back-of-envelope calculation

Why ~ 40K traces?

Most edges are incident on nodes having small degree



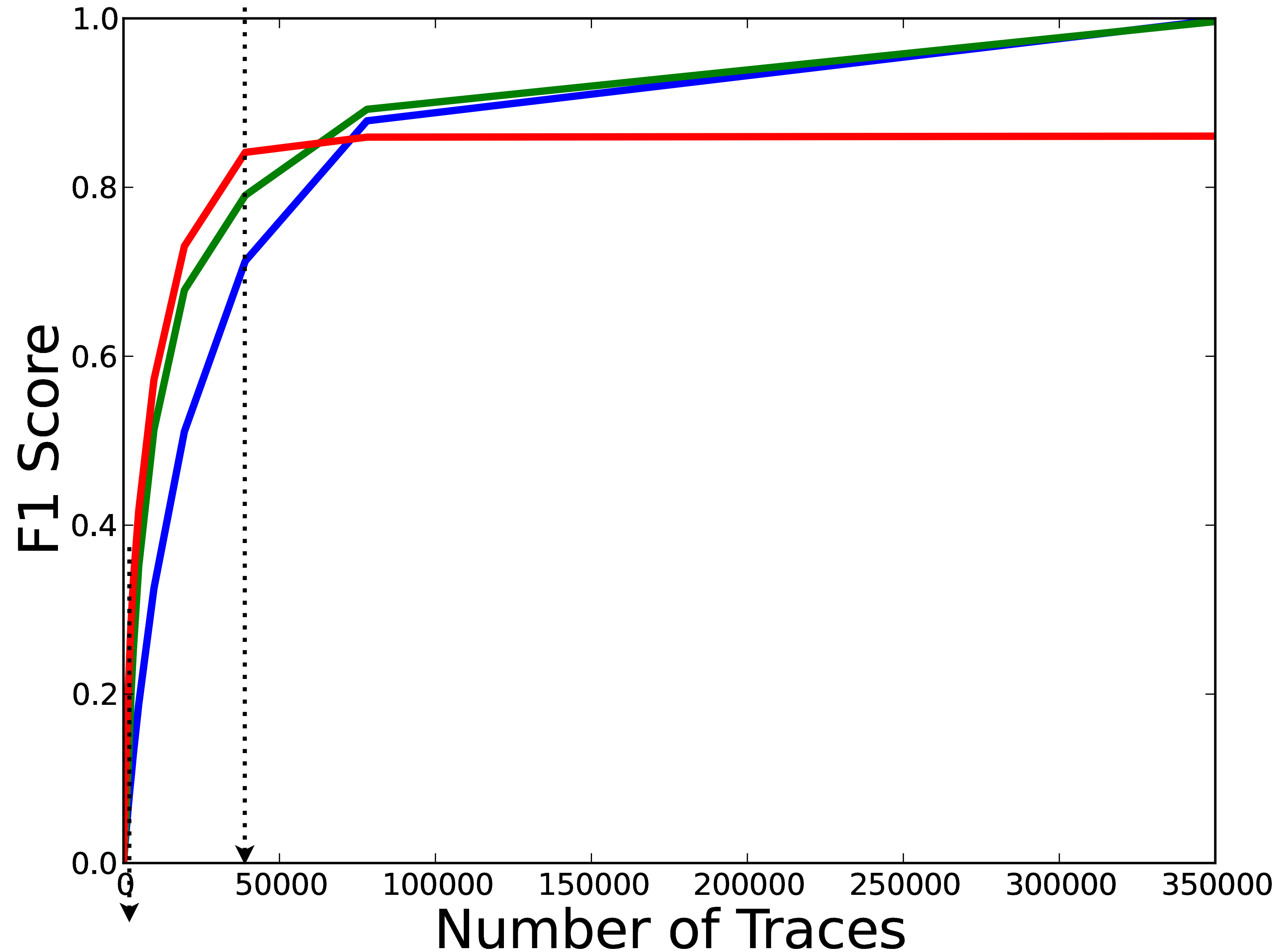
Facebook-Rice Undergrad

Back-of-envelope calculation

The graph has 1220 nodes and average degree equal to 35.41…

Why ~ 40K traces?

Most edges are incident on nodes having small degree



Facebook-Rice Undergrad

Back-of-envelope calculation

The graph has 1220 nodes and average degree equal to 35.41…

1220 x 35.41 = 43200.2

Too many traces?

- In some cases such a large number of traces is not available.

- In those cases, one can still try to reconstruct some (less precise) missing information about the network.

E-mail Activisim

Joint with

Jon Kleinberg, David Liben-Nowell

Internet Activism

- Very important phenomenon
- Incomplete Traces
- Chain Letter Petitions:
  how to estimate the reach?

NPR Chain Letter

PBS, NPR (National Public Radio), and the arts are facing major cutbacks in funding. In spite of the efforts of each station to reduce spending costs and streamline their services, the government officials believe that the funding currently going to these programs is too large a portion of funding for something which is seen as "unworthwhile."

[...]

When this issue comes up in 1996, the funding will be determined for fiscal years 1996-1998.

The only way that our representatives can be aware of the base of support or PBS and funding for these types of programs is by making our voices heard.

Please add your name to this list if you believe in what we stand for. This list will be forwarded to the President of the United States, the Vice President of the United States, the House of Representatives and Congress.

If you happen to be the 50th, 100th, 150th, etc. signer of this petition, please forward to: kubi7975@blue.univnorthco.edu . This way we can keep track of the lists and organize them. Forward this to everyone you know, and help us to keep these programs alive.

Thank you.

1. Elizabeth Weinert, student, University of Northern Colorado, Greeley, Colorado.
2. Robert M. Penn; San Francisco, CA
3. Gregory S. Williamson, San Francisco, CA
4. Daniel C. Knightly, Austin, TX
5. Andrew H. Knightly, Los Angeles, CA
6. Aaron C. Yeater, Somerville, MA
7. Tobie M. Cornejo, Washington, DC
8. John T. Mason, Dalton, MA
9. Eric W. Fish, Williamstown, MA
10. Courtney E. Estill, Hamilton College, NY
11. Vanessa Moore, Northfield, MN
12. Lynne Raschke, Haverford College, PA (originally Minnesota)
13. Deborah Bielak, Haverford, PA
14. Morgan Lloyd, Haverford, PA 19041
15. Galen Lloyd, Goucher College, MD
16. Brian Eastwood, University of Vermont, VT
17. Elif Batuman, Harvard University, MA
18. Kohar Jones, Yale University, CT
19. Claudia Brittenham, Yale University, CT
20. Alexandra Block, Yale University, CT
21. Susanna Chu, Yale University, CT
22. Michelle Chen, Harvard University, MA
23. Jessica Hammer, Harvard University, MA
24. Ann Pettigrew, Haverford College, PA
25. Kirstin Knox, Swarthmore College, PA
26. Jason Adler, Swarthmore College, PA
27. Daniel Gottlieb, Swarthmore College (but truly from Lawrence, KS)
28. Josh Feltman, Tufts University, MA
29. Louise Forrest, Massachusetts Institute of Technology, MA
30. HongSup Park, Massachusetts Institute of Technology, MA (originally from Portage, Wisconsin)
31. Ana Sandoval,Massachusetts Institute of Technology

[...]

NPR Chain Letter

PBS, NPR (National Public Radio), and the arts are facing major cutbacks in funding. In spite of the efforts of each station to reduce spending costs and streamline their services, the government officials believe that the funding currently going to these programs is too large a portion of funding for something which is seen as "unworthwhile."

[...]

When this issue comes up in 1996, the funding will be determined for fiscal years 1996-1998.

The only way that our representatives can be aware of the base of support or PBS and funding for these types of programs is by making our voices heard.

Please add your name to this list if you believe in what we stand for. This list will be forwarded to the President of the United States, the Vice President of the United States, the House of Representatives and Congress.

If you happen to be the 50th, 100th, 150th, etc. signer of this petition, please forward to: kubi7975@blue.univnorthco.edu . This way we can keep track of the lists and organize them. *Forward this to everyone you know, and help us to keep these programs alive.*

Thank you.

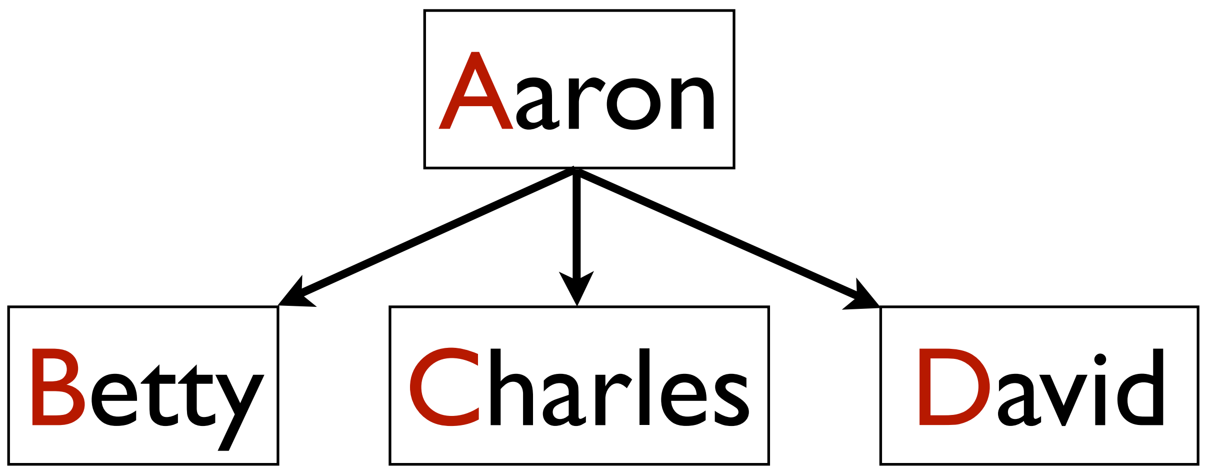
1. Elizabeth Weinert, student, University of Northern Colorado, Greeley, Colorado.
2. Robert M. Penn; San Francisco, CA
3. Gregory S. Williamson, San Francisco, CA
4. Daniel C. Knightly, Austin, TX
5. Andrew H. Knightly, Los Angeles, CA
6. Aaron C. Yeater, Somerville, MA
7. Tobie M. Cornejo, Washington, DC
8. John T. Mason, Dalton, MA
9. Eric W. Fish, Williamstown, MA
10. Courtney E. Estill, Hamilton College, NY
11. Vanessa Moore, Northfield, MN
12. Lynne Raschke, Haverford College, PA (originally Minnesota)
13. Deborah Bielak, Haverford, PA
14. Morgan Lloyd, Haverford, PA 19041
15. Galen Lloyd, Goucher College, MD
16. Brian Eastwood, University of Vermont, VT
17. Elif Batuman, Harvard University, MA
18. Kohar Jones, Yale University, CT
19. Claudia Brittenham, Yale University, CT
20. Alexandra Block, Yale University, CT
21. Susanna Chu, Yale University, CT
22. Michelle Chen, Harvard University, MA
23. Jessica Hammer, Harvard University, MA
24. Ann Pettigrew, Haverford College, PA
25. Kirstin Knox, Swarthmore College, PA
26. Jason Adler, Swarthmore College, PA
27. Daniel Gottlieb, Swarthmore College (but truly from Lawrence, KS)
28. Josh Feltman, Tufts University, MA
29. Louise Forrest, Massachusetts Institute of Technology, MA
30. HongSup Park, Massachusetts Institute of Technology, MA (originally from Portage, Wisconsin)
31. Ana Sandoval,Massachusetts Institute of Technology

[...]

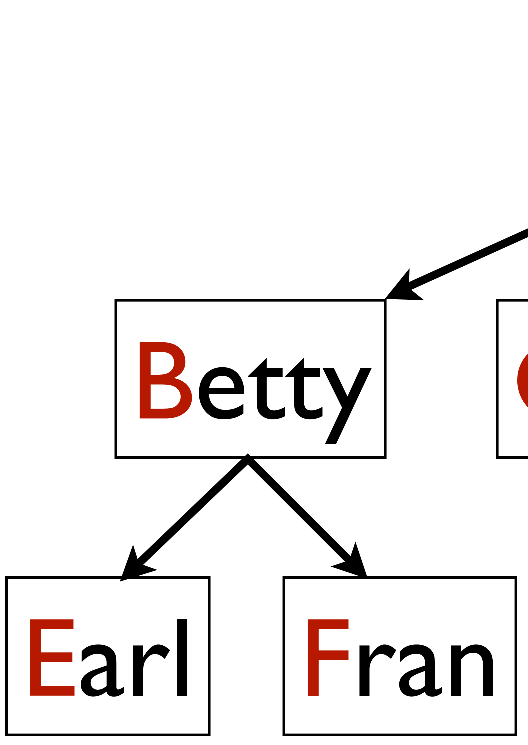
Chain Letters

Aaron

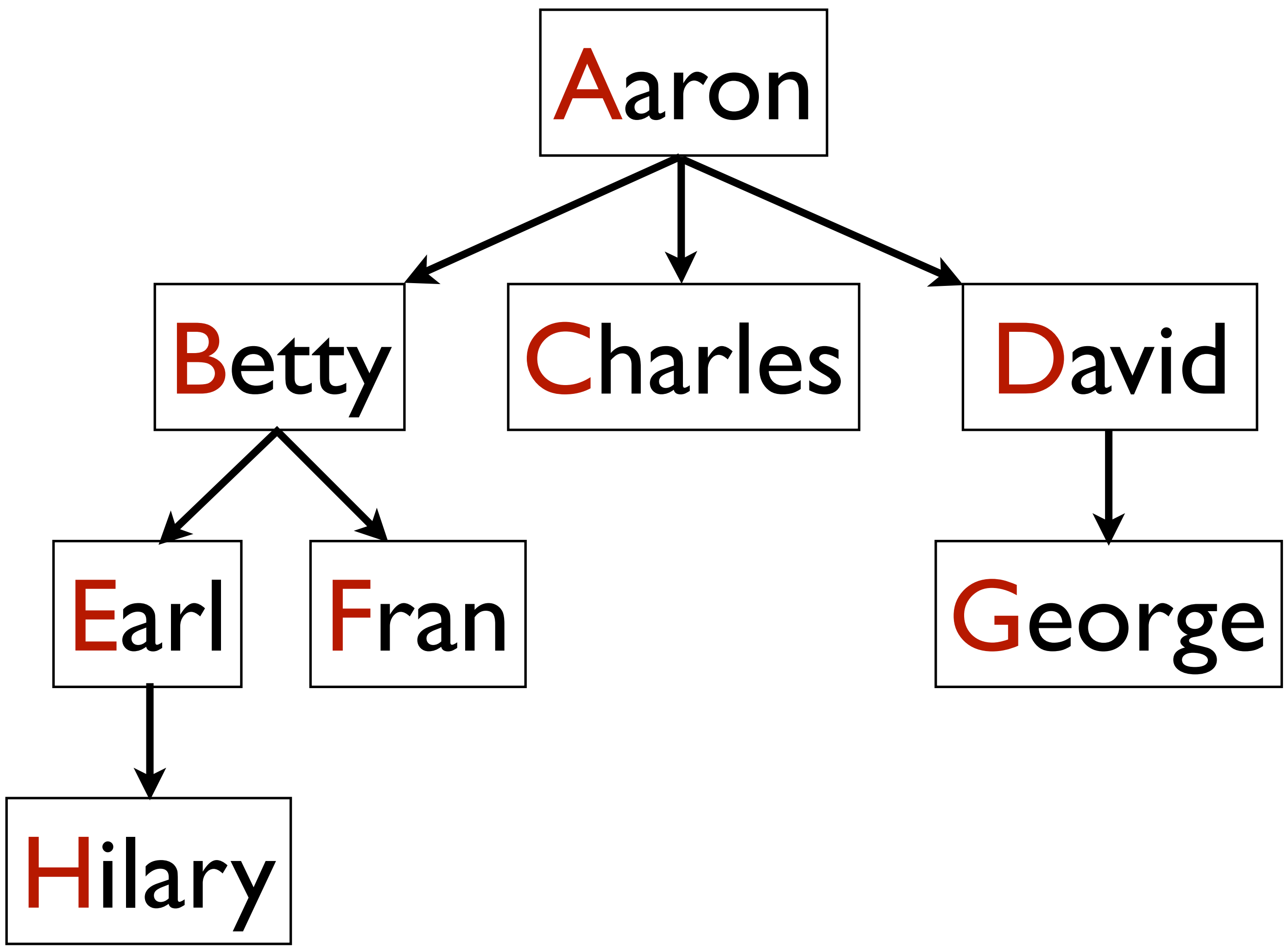
Chain Letters



Chain Letters

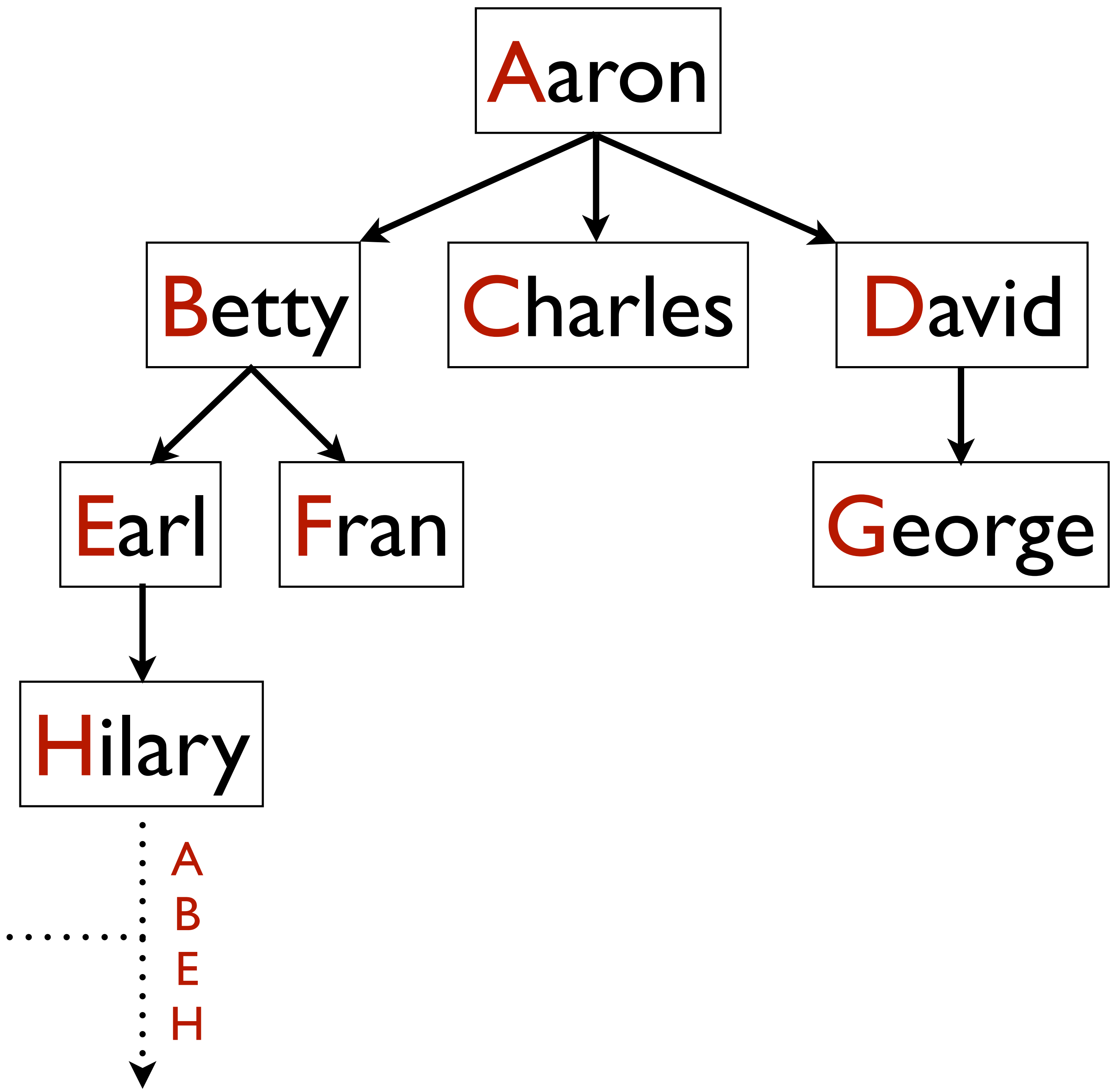


Chain Letters

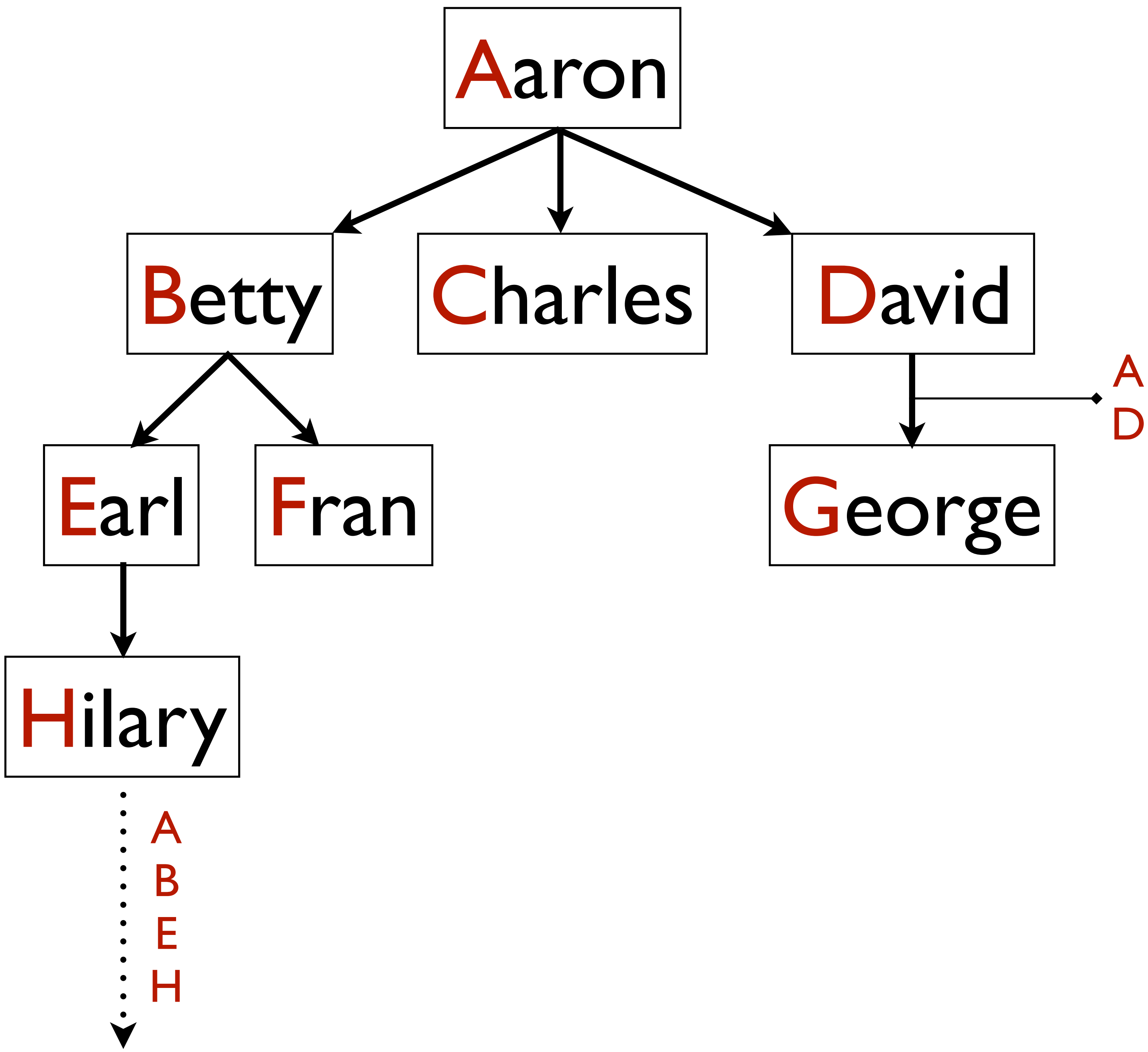


Chain Letters

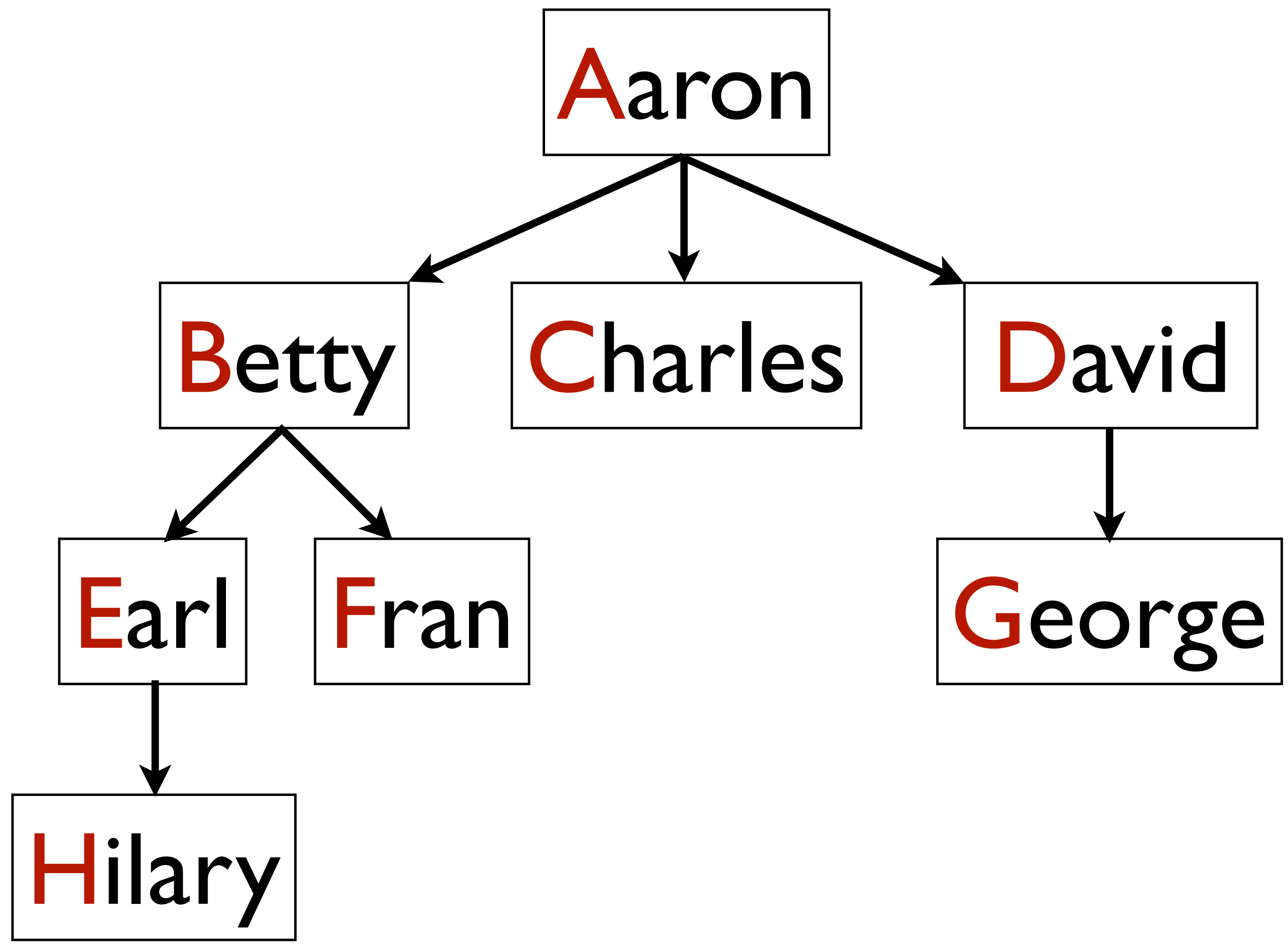
Dear all,
an important
cause demands
your attention.
[…]
If you care
about this, add
your name and
forward this
letter.
[…]
The signers,
Aaron
Betty
Earl
Hilary



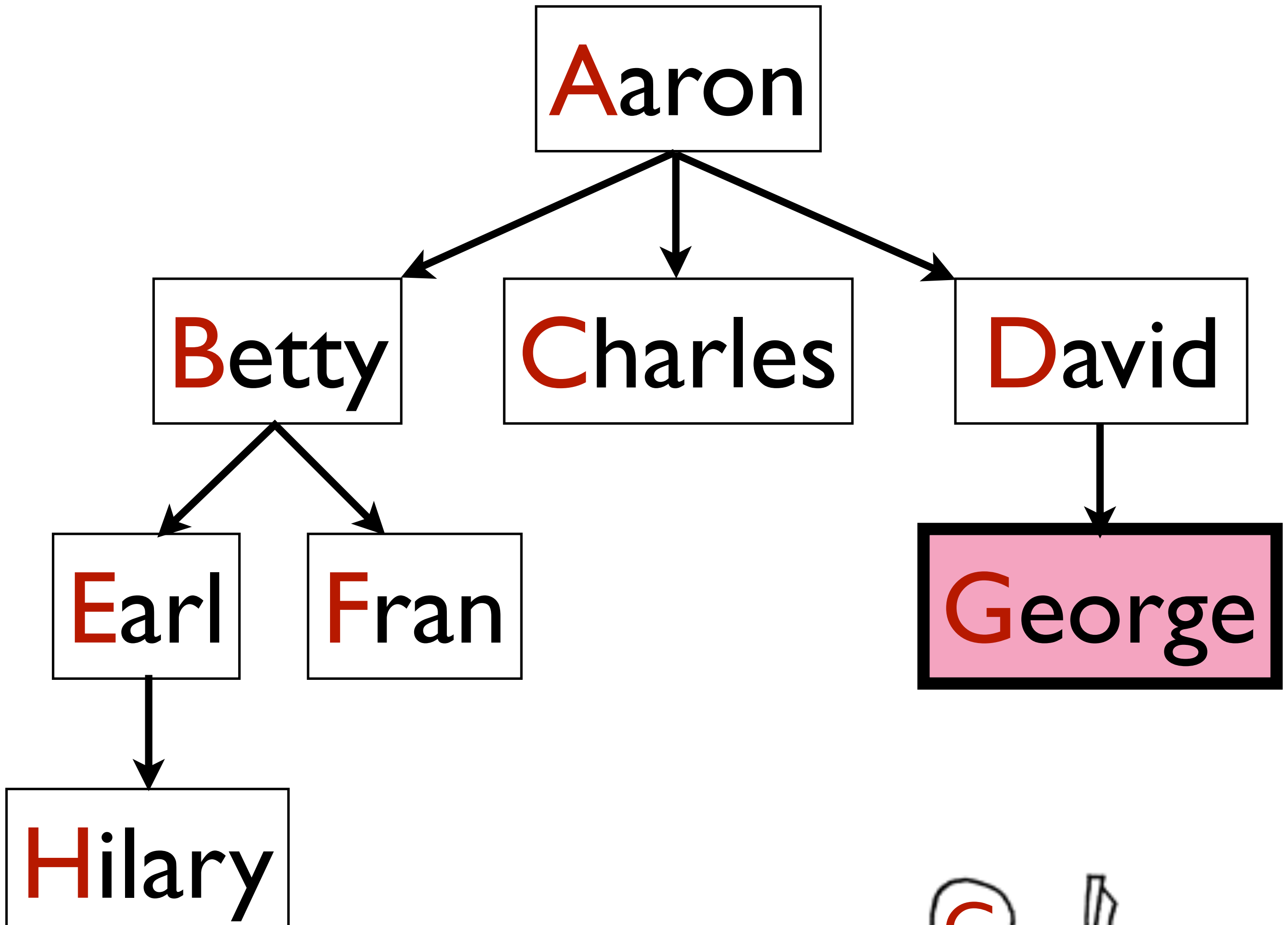
Chain Letters



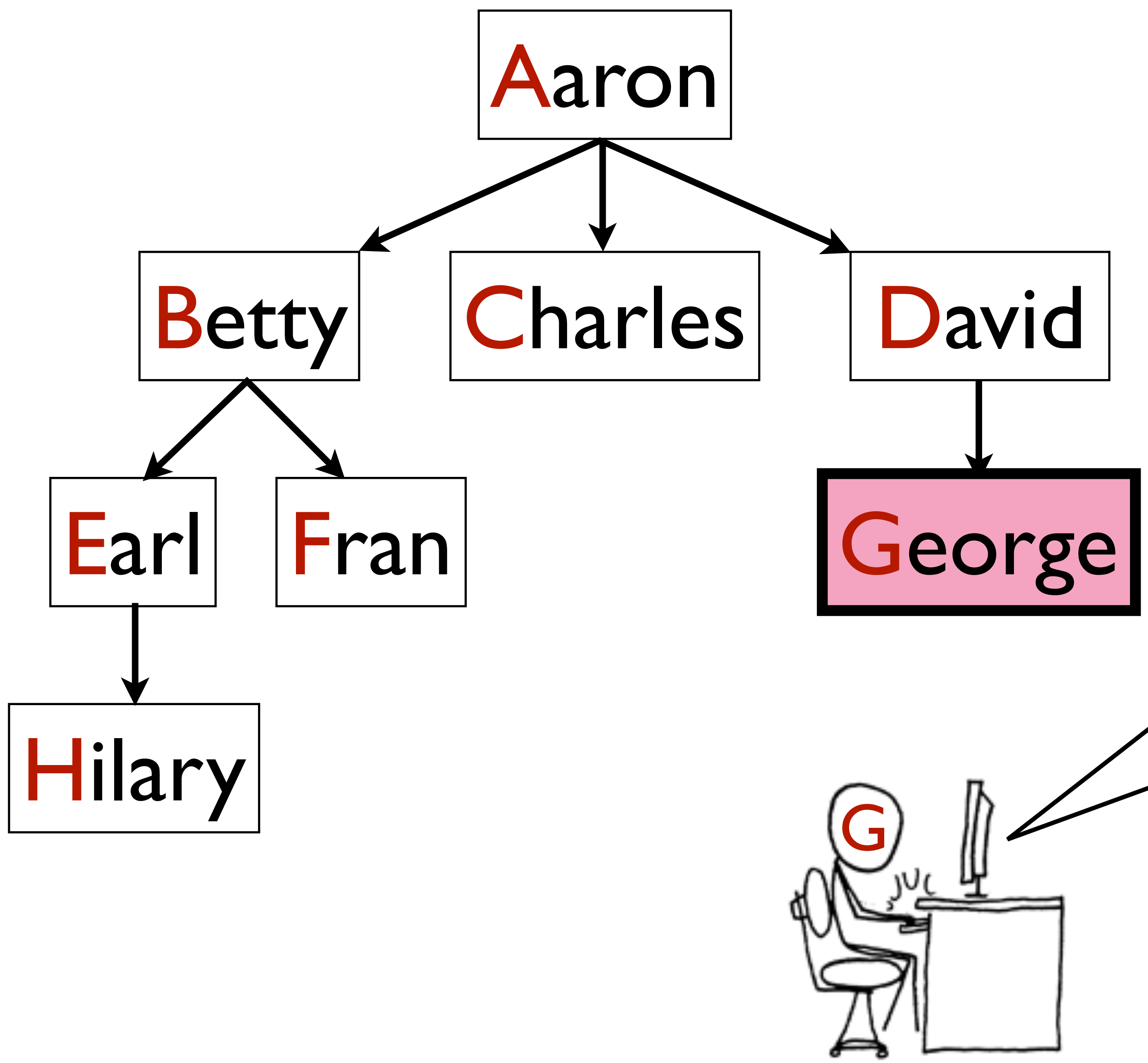
Chain Letters



Chain Letters



Chain Letters

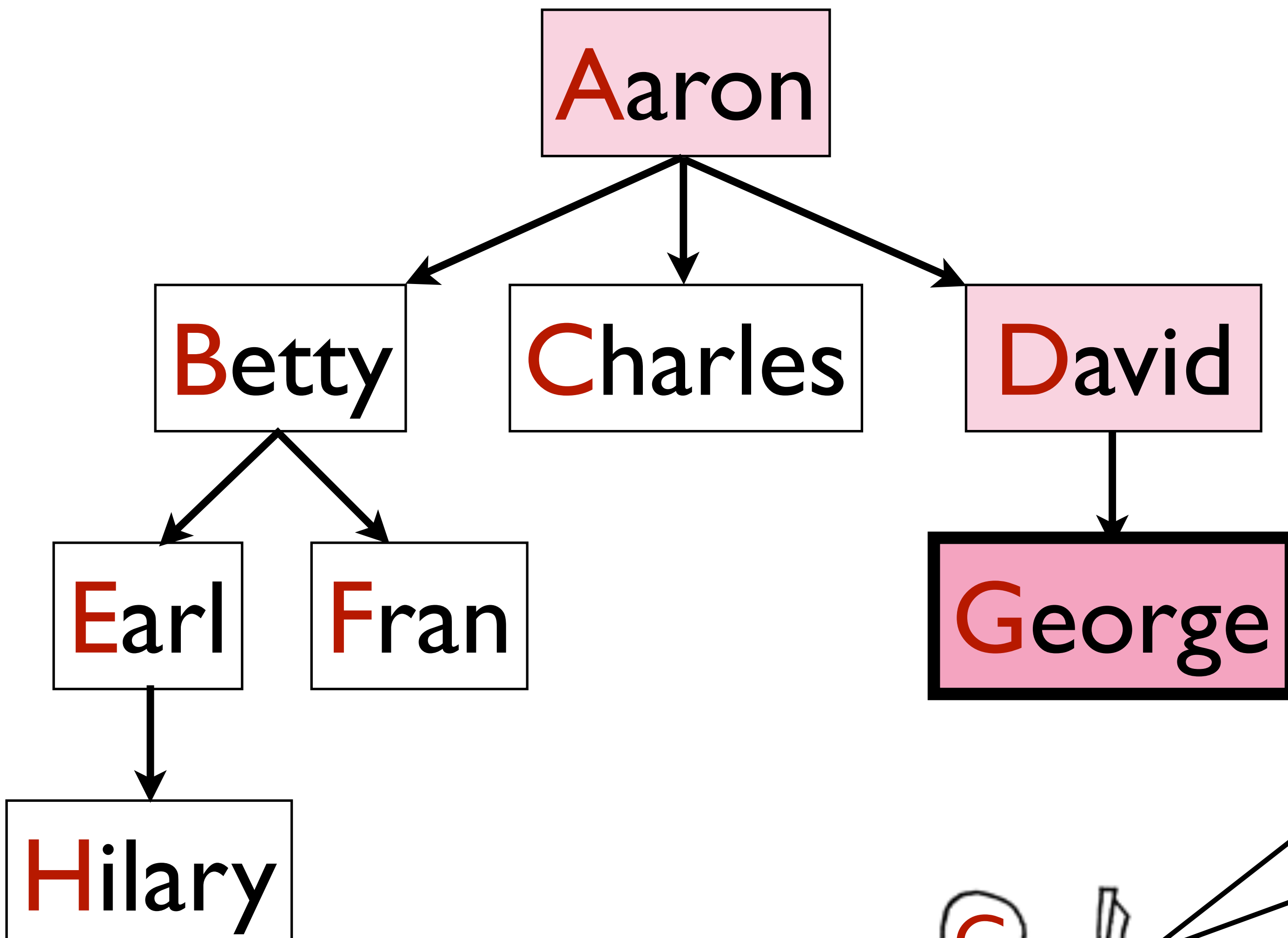


George's Blog

Here is something that I sent to my friends today:

```
Dear all,
an important
cause demands
your attention.
[…]
If you care about
this, add your
name and forward
this letter.
[…]
The signers,
Aaron
David
George
```

Chain Letters

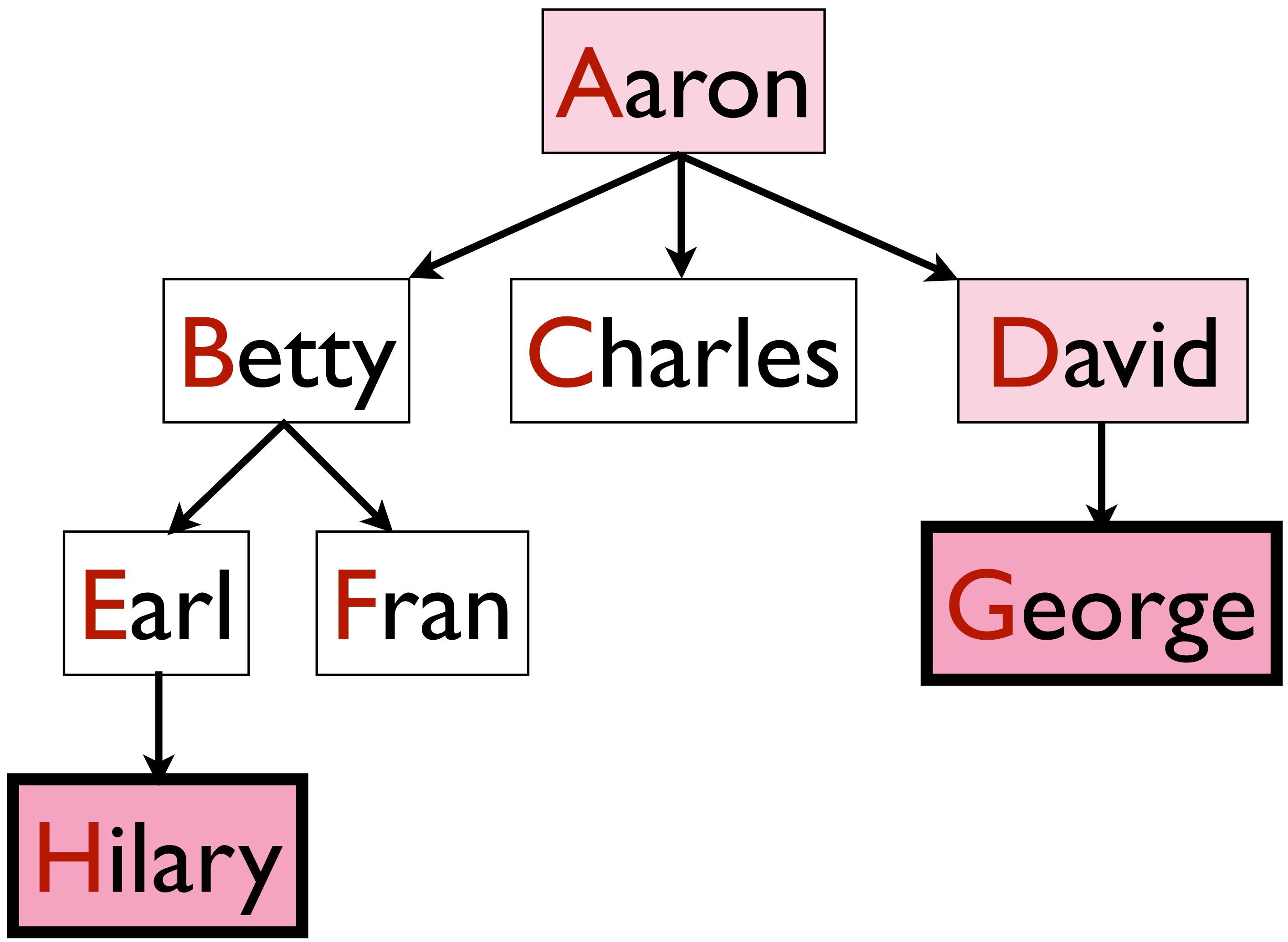


George's Blog

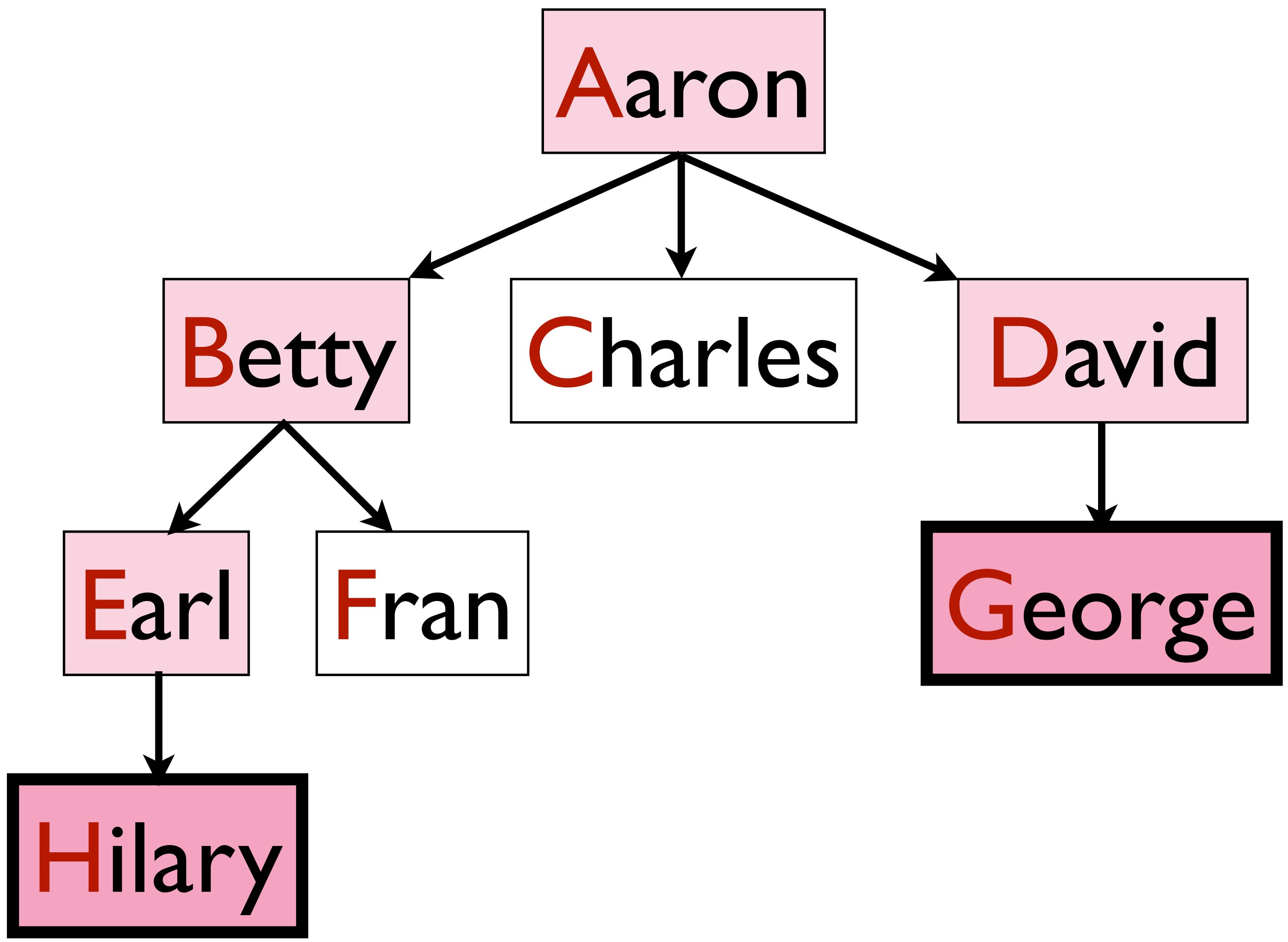
Here is something that I sent to my friends today:

```
Dear all,
an important
cause demands
your attention.
[…]
If you care about
this, add your
name and forward
this letter.
[…]
The signers,
Aaron
David
George
```

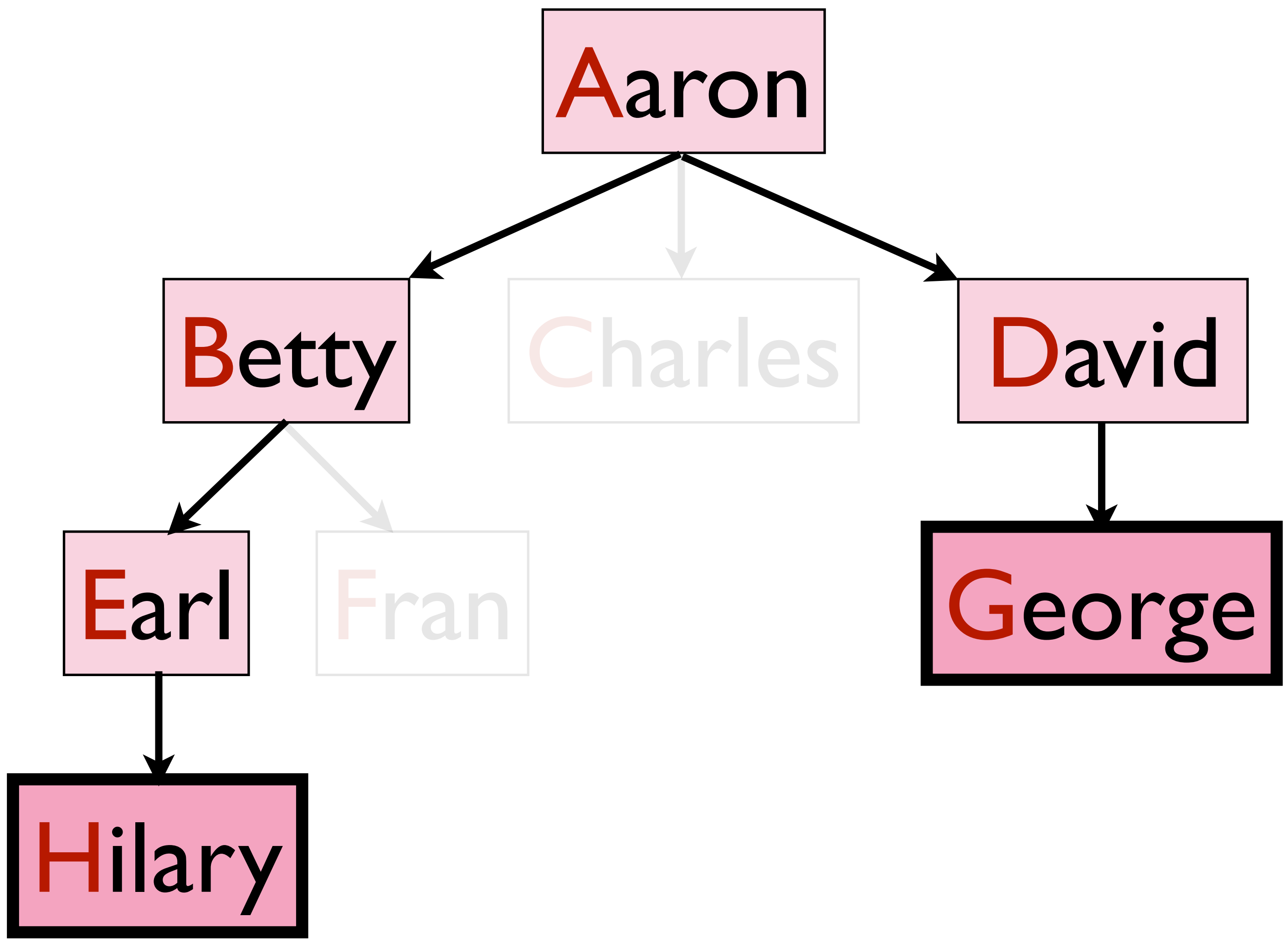
Chain Letters



Chain Letters



Chain Letters



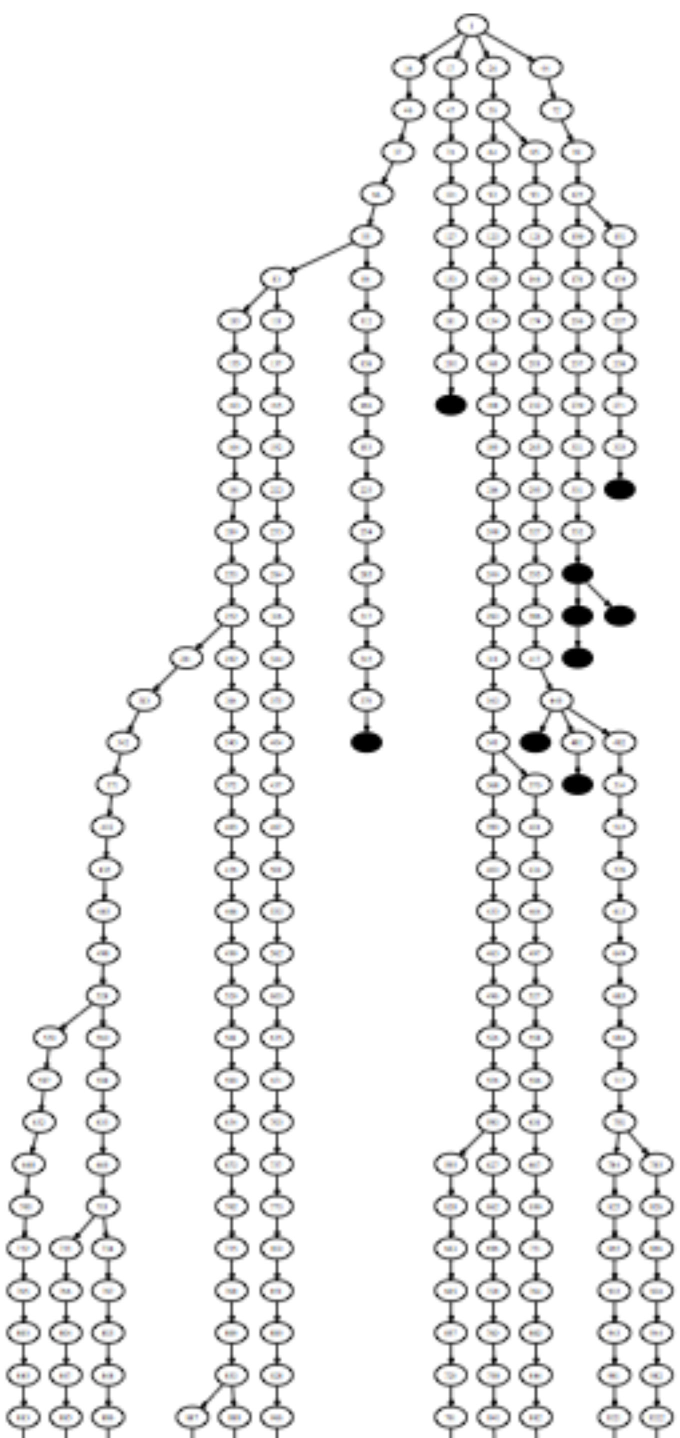
George and Hilary, by *exposing* their emails, *revealed* a subtree of the Chain Letter tree.

Real-World Chain Letters' Tree

- *[Liben-Nowell, Kleinberg, PNAS'08]*, mined
  - web-accessible mailing-lists,
  - blog posts.
- They obtained some *"exposed"* nodes of two Chain Letters' trees, and
- they produced two *"revealed"* trees.

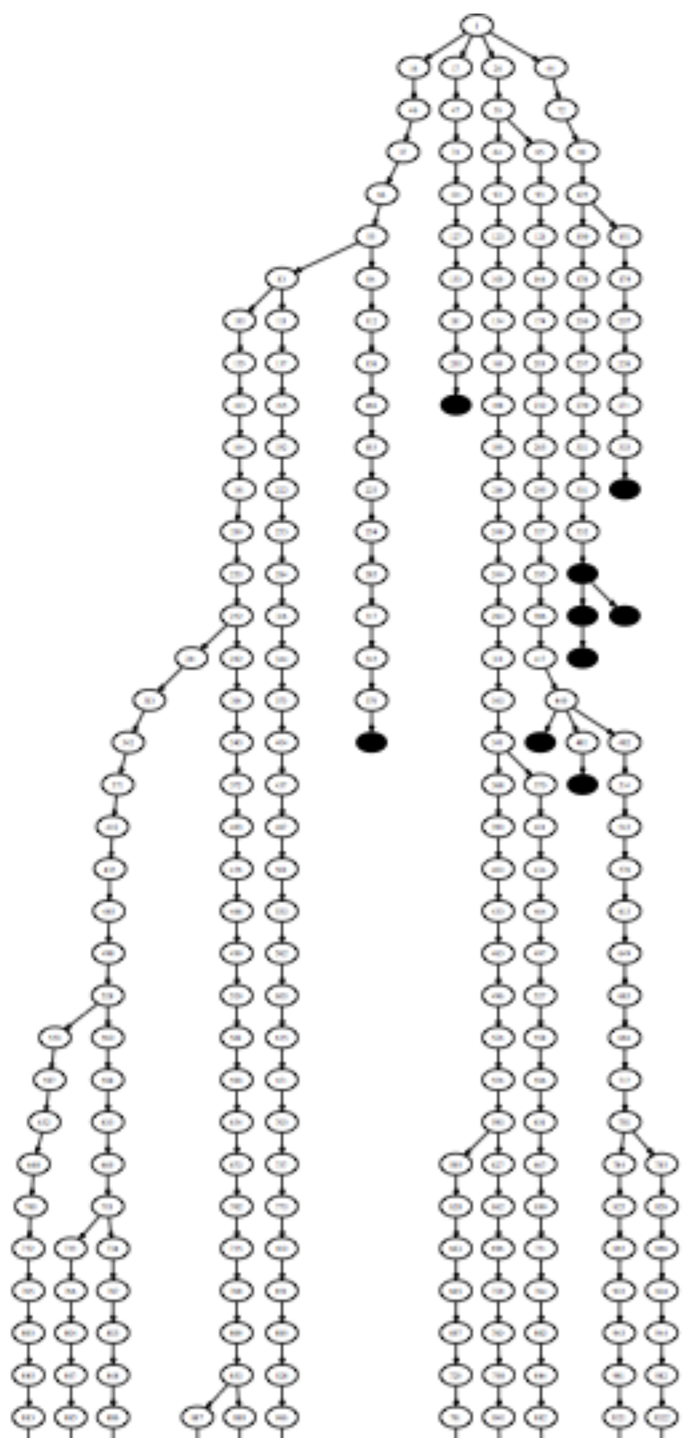
NPR revealed tree

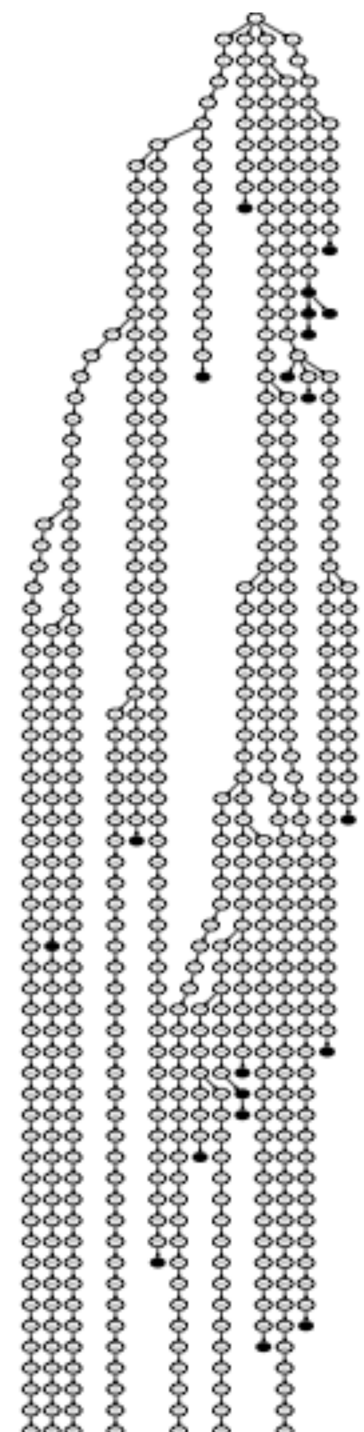
Liben-Nowell, Kleinberg, PNAS'08



NPR revealed tree

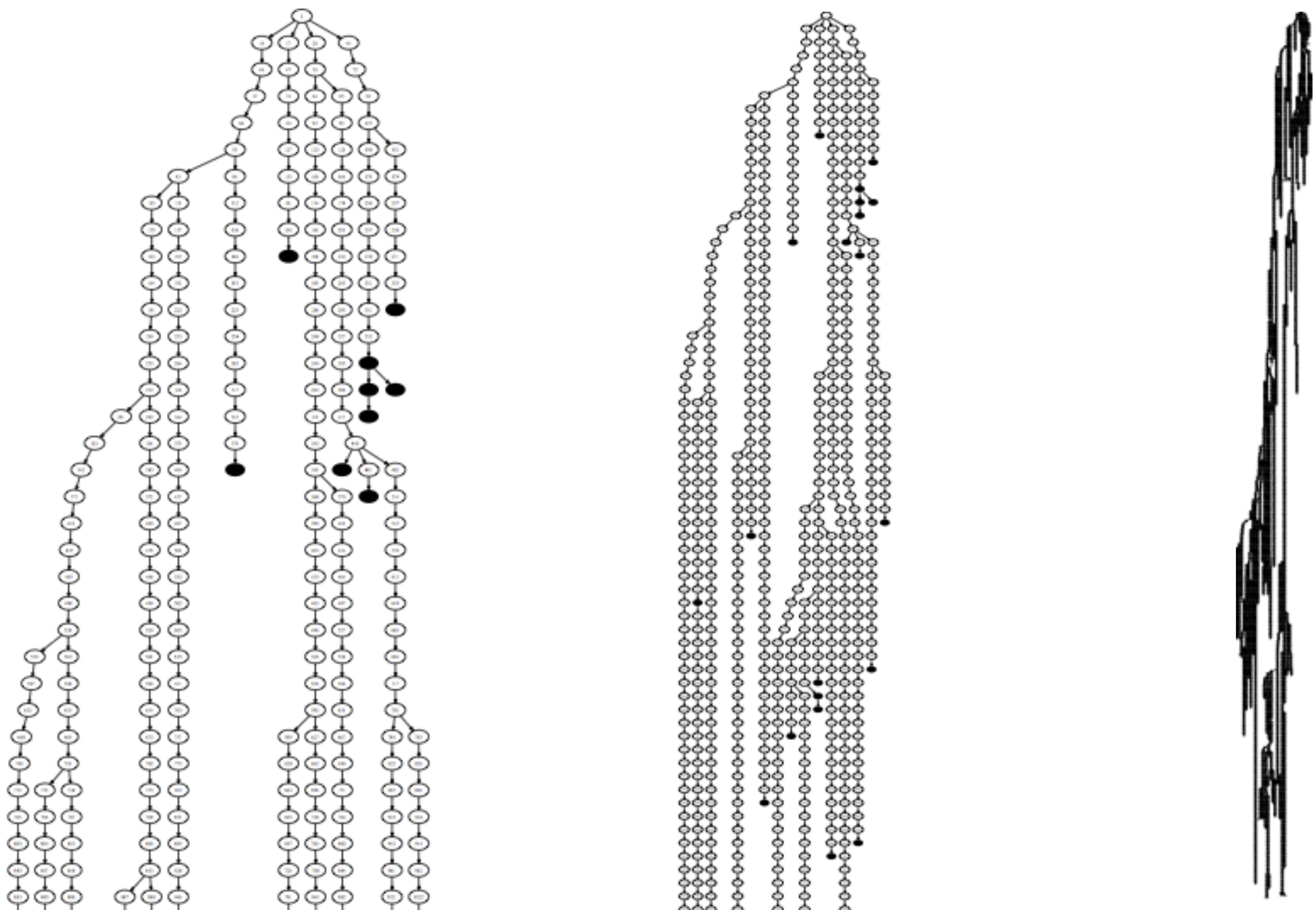
Liben-Nowell, Kleinberg, PNAS'08





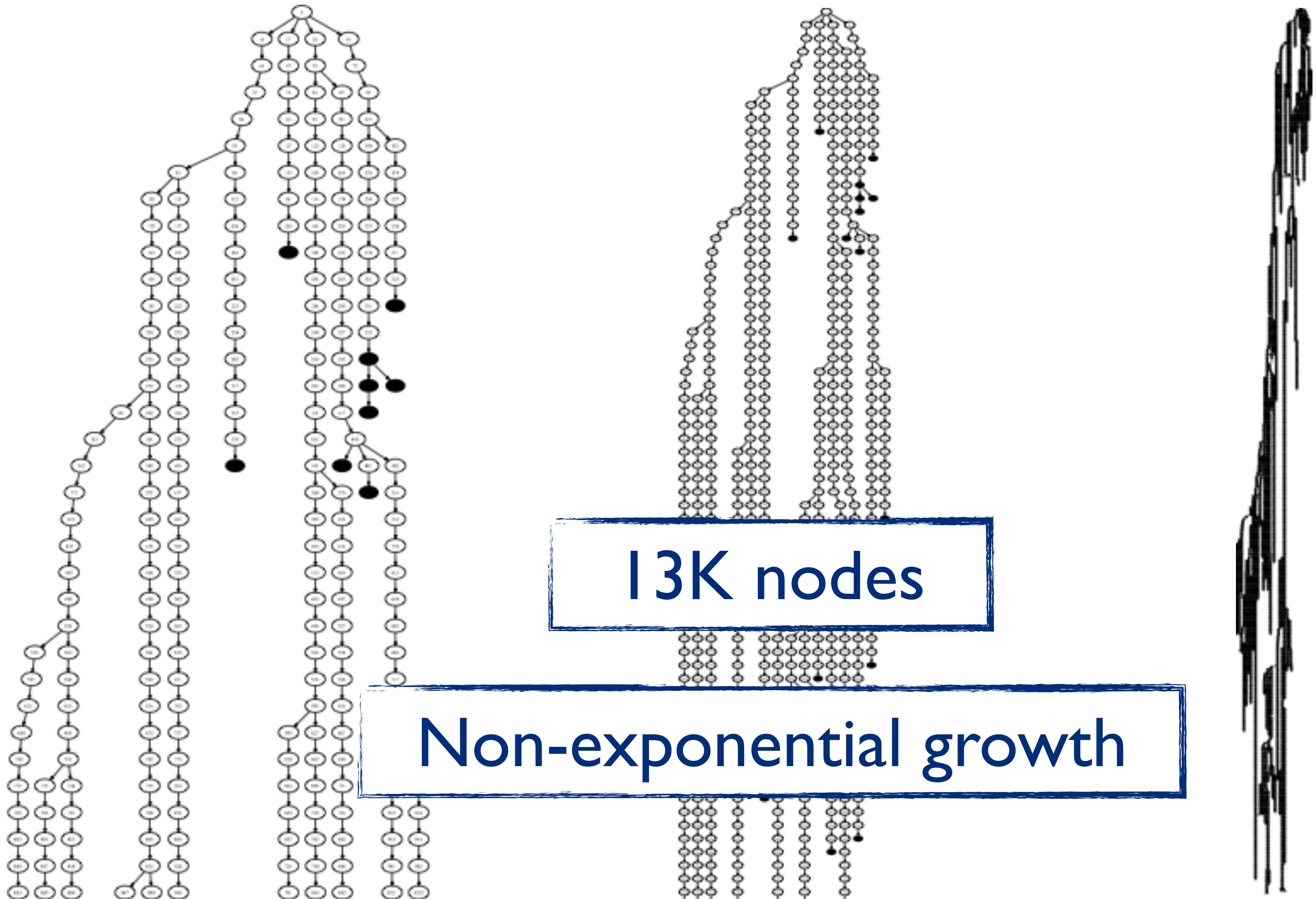
NPR revealed tree

Liben-Nowell, Kleinberg, PNAS'08



NPR revealed tree

Liben-Nowell, Kleinberg, PNAS'08



13K nodes

Non-exponential growth

Iraq Chain Letter

Dear all:

The US Congress has just authorized the President of the US to go to war against Iraq. The UN is gathering signatures in an effort to avoid this tragic world event.

Please consider this an urgent request: UN Petition for Peace - Stand for Peace. Islam is not the Enemy.
War is NOT the Answer.

Today we are at a point of imbalance in the world and are moving toward what may be the beginning of a THIRD WORLD WAR.

Please COPY (rather than Forward) this e-mail in a new message, sign at the end of the list, and send it to all the people whom you know.

If you receive this list with more than 500 names signed, please send a copy of the message to:

usa@un.int
president@whitehouse.gov

Even if you decide not to sign, please consider forwarding the petition
on instead of
deleting it.

1) Suzanne Dathe, Grenoble, France
2) Laurence COMPARAT, Grenoble, France
3) Philippe MOTTE, Grenoble, France
4) Jok FERRAND, Mont St. Martin, France
5) Emmanuelle PIGNOL, St Martin d'Heres, FRANCE
6) Marie GAUTHIER, Grenoble, FRANCE
7) Laurent VESCALO, Grenoble, FRANCE
8) Mathieu MOY, St Egreve, FRANCE
9) Bernard BLANCHET, Mont St Martin,FRANCE
10) Tassadite FAVRIE, Grenoble, FRANCE
11) Loic GODARD, St Ismier, FRANCE
12) Benedicte PASCAL, Grenoble, FRANCE
13) Khedaidja BENATIA, Grenoble, FRANCE
14) Marie-Therese LLORET, Grenoble,FRANCE
15) Benoit THEAU, Poitiers, FRANCE
16) Bruno CONSTANTIN, Poitiers, FRANCE
17) Christian COGNARD, Poitiers, FRANCE
18) Robert GARDETTE, Paris, FRANCE
19) Claude CHEVILLARD, Montpellier, FRANCE
20) Gilles FREISS, Montpellier, FRANCE
21) Patrick AUGEREAU, Montpellier, FRANCE
22) Jean IMBER! T, Marseille, FRANCE
23) Jean-Claude MURAT, Toulouse, France
24) Anna BASSOLS, Barcelona, Catalonia
25) Mireia DUNACH, Barcelona, Catalonia
26) Michel VILLAZ, Grenoble, France
27) Pages Frederique, Dijon, France
28) Rodolphe FISCHMEISTER,Chatenay-Malabry, France
29) Francois BOUTEAU, Paris, France
30) Patrick PETER, Paris, France
31) Lorenza RADICI, Paris, France
32) Monika Siegenthaler, Bern, Switzerland
33) Mark Philp,Glasgow,Scotland
34) Tomas Andersson, Stockholm, Sweden
35) Jonas Eriksson, Stockholm, Sweden
36) Karin Eriksson, Stockholm, Sweden
...

Iraq Chain Letter

Dear all:

The US Congress has just authorized the President of the US to go to war against Iraq. The UN is gathering signatures in an effort to avoid this tragic world event.

Please consider this an urgent request: UN Petition for Peace - Stand for Peace. Islam is not the Enemy.
War is NOT the Answer.

Today we are at a point of imbalance in the world and are moving toward what may be the beginning of a THIRD WORLD WAR.

Please COPY (rather than Forward) this e-mail in a new message, sign at the end of the list, and send it to all the people whom you know.

If you receive this list with more than 500 names signed, please send a copy of the message to:

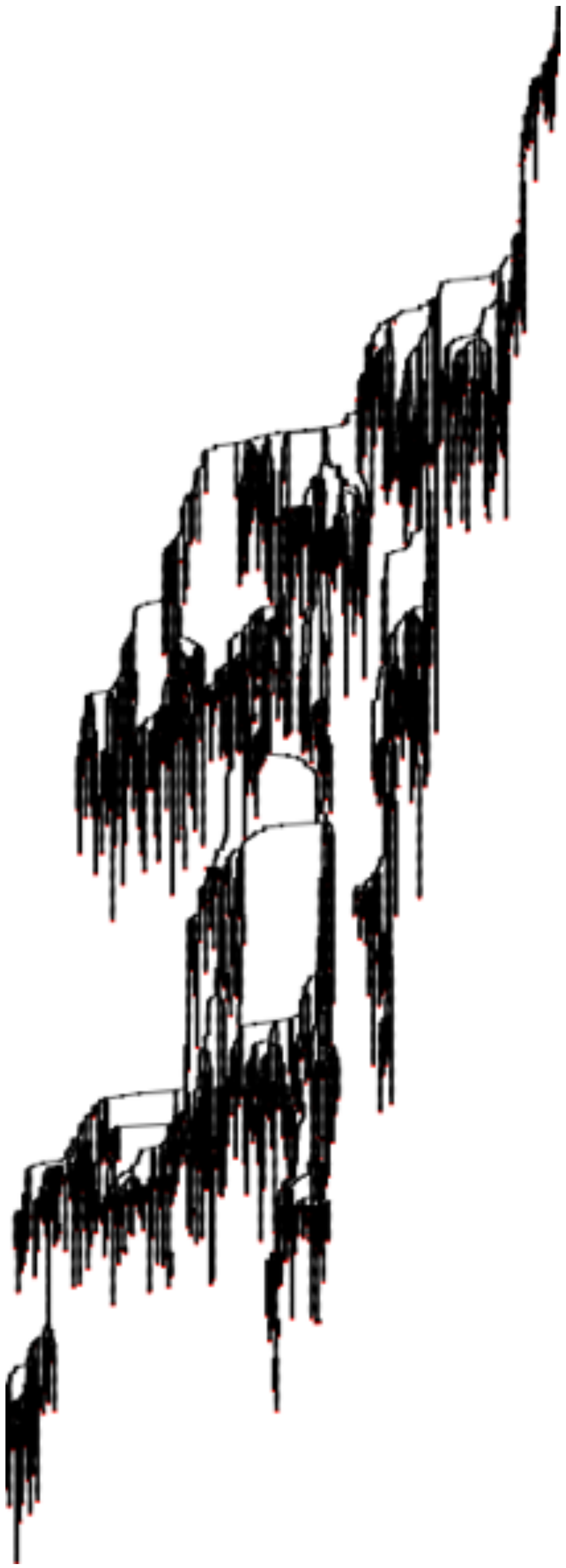
usa@un.int
president@whitehouse.gov

Even if you decide not to sign, please consider forwarding the petition
on instead of
deleting it.

1) Suzanne Dathe, Grenoble, France
2) Laurence COMPARAT, Grenoble, France
3) Philippe MOTTE, Grenoble, France
4) Jok FERRAND, Mont St. Martin, France
5) Emmanuelle PIGNOL, St Martin d'Heres, FRANCE
6) Marie GAUTHIER, Grenoble, FRANCE
7) Laurent VESCALO, Grenoble, FRANCE
8) Mathieu MOY, St Egreve, FRANCE
9) Bernard BLANCHET, Mont St Martin,FRANCE
10) Tassadite FAVRIE, Grenoble, FRANCE
11) Loic GODARD, St Ismier, FRANCE
12) Benedicte PASCAL, Grenoble, FRANCE
13) Khedaidja BENATIA, Grenoble, FRANCE
14) Marie-Therese LLORET, Grenoble,FRANCE
15) Benoit THEAU, Poitiers, FRANCE
16) Bruno CONSTANTIN, Poitiers, FRANCE
17) Christian COGNARD, Poitiers, FRANCE
18) Robert GARDETTE, Paris, FRANCE
19) Claude CHEVILLARD, Montpellier, FRANCE
20) Gilles FREISS, Montpellier, FRANCE
21) Patrick AUGEREAU, Montpellier, FRANCE
22) Jean IMBER! T, Marseille, FRANCE
23) Jean-Claude MURAT, Toulouse, France
24) Anna BASSOLS, Barcelona, Catalonia
25) Mireia DUNACH, Barcelona, Catalonia
26) Michel VILLAZ, Grenoble, France
27) Pages Frederique, Dijon, France
28) Rodolphe FISCHMEISTER,Chatenay-Malabry, France
29) Francois BOUTEAU, Paris, France
30) Patrick PETER, Paris, France
31) Lorenza RADICI, Paris, France
32) Monika Siegenthaler, Bern, Switzerland
33) Mark Philp,Glasgow,Scotland
34) Tomas Andersson, Stockholm, Sweden
35) Jonas Eriksson, Stockholm, Sweden
36) Karin Eriksson, Stockholm, Sweden
...

IRAQ revealed tree

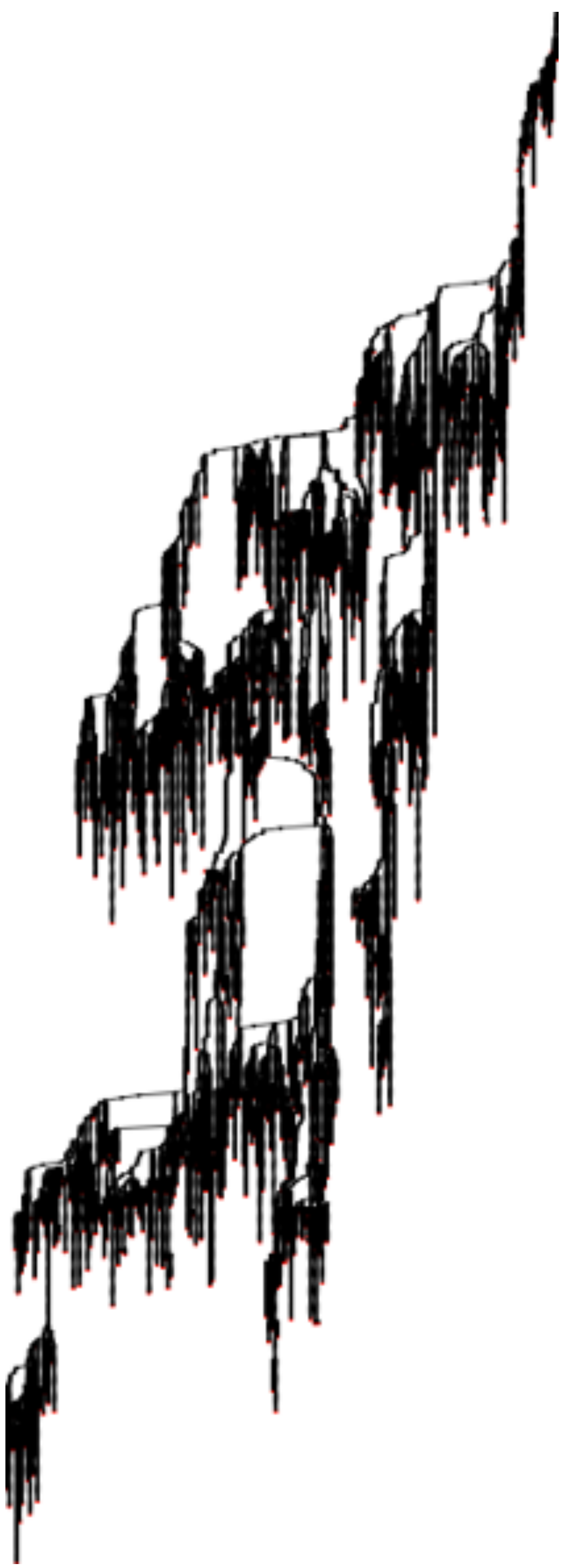
Liben-Nowell, Kleinberg, PNAS'08



18,119 nodes

IRAQ revealed tree

Liben-Nowell, Kleinberg, PNAS'08

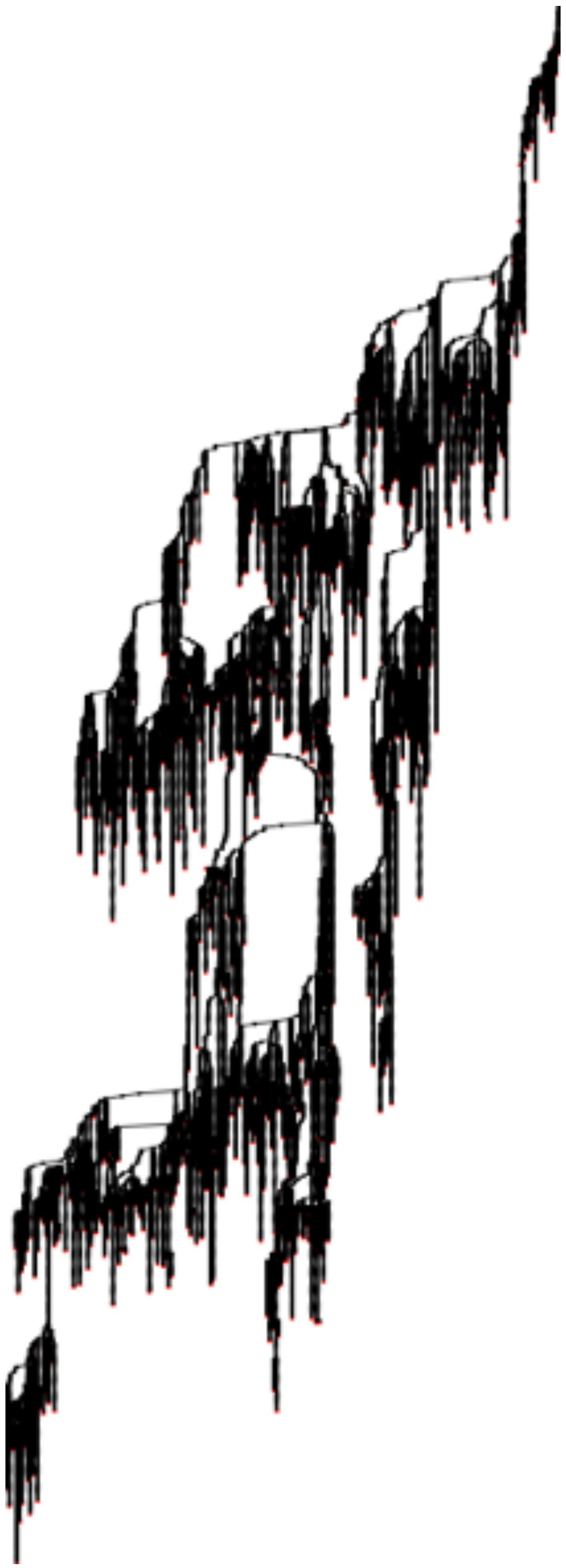


18,119 nodes

17,079 nodes with one child *(94%)*

IRAQ revealed tree

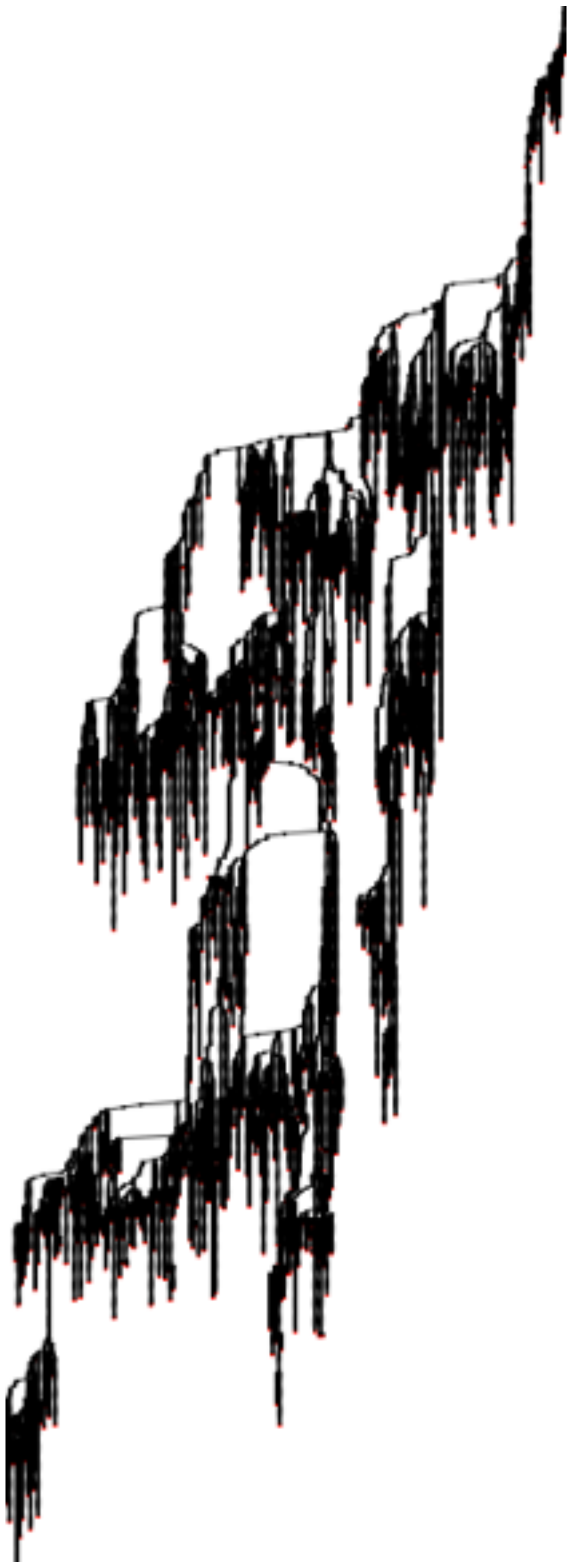
Liben-Nowell, Kleinberg, PNAS'08



18,119 nodes
17,079 nodes with one child *(94%)*
620 exposed nodes
557 (exposed) leaves

IRAQ revealed tree

Liben-Nowell, Kleinberg, PNAS'08



18,119 nodes

17,079 nodes with one child *(94%)*

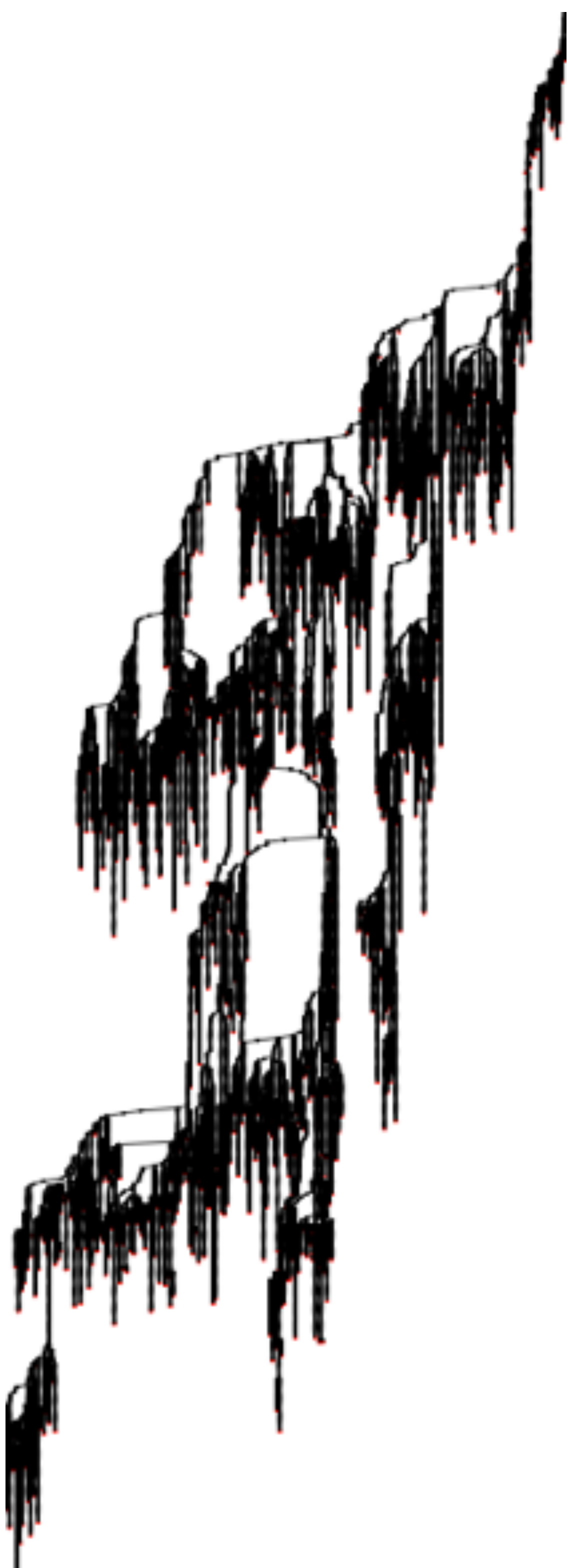
620 exposed nodes

557 (exposed) leaves

Why is this fraction so high?

IRAQ revealed tree

Liben-Nowell, Kleinberg, PNAS'08



18,119 nodes

17,079 nodes with one child *(94%)*

620 exposed nodes

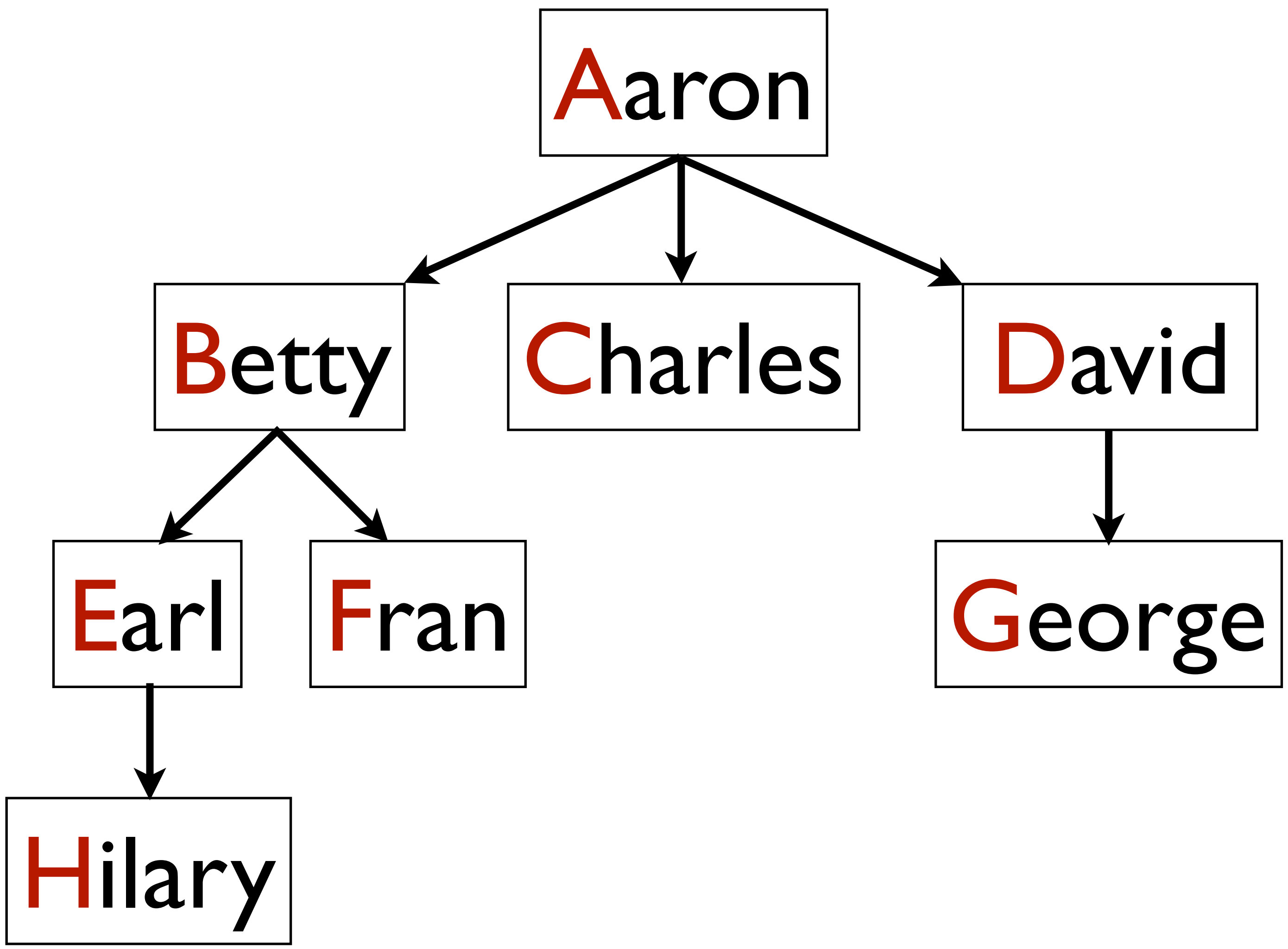
557 (exposed) leaves

Why is this fraction so high?

What can we infer about the original, *unknown*, Chain Letter Tree?

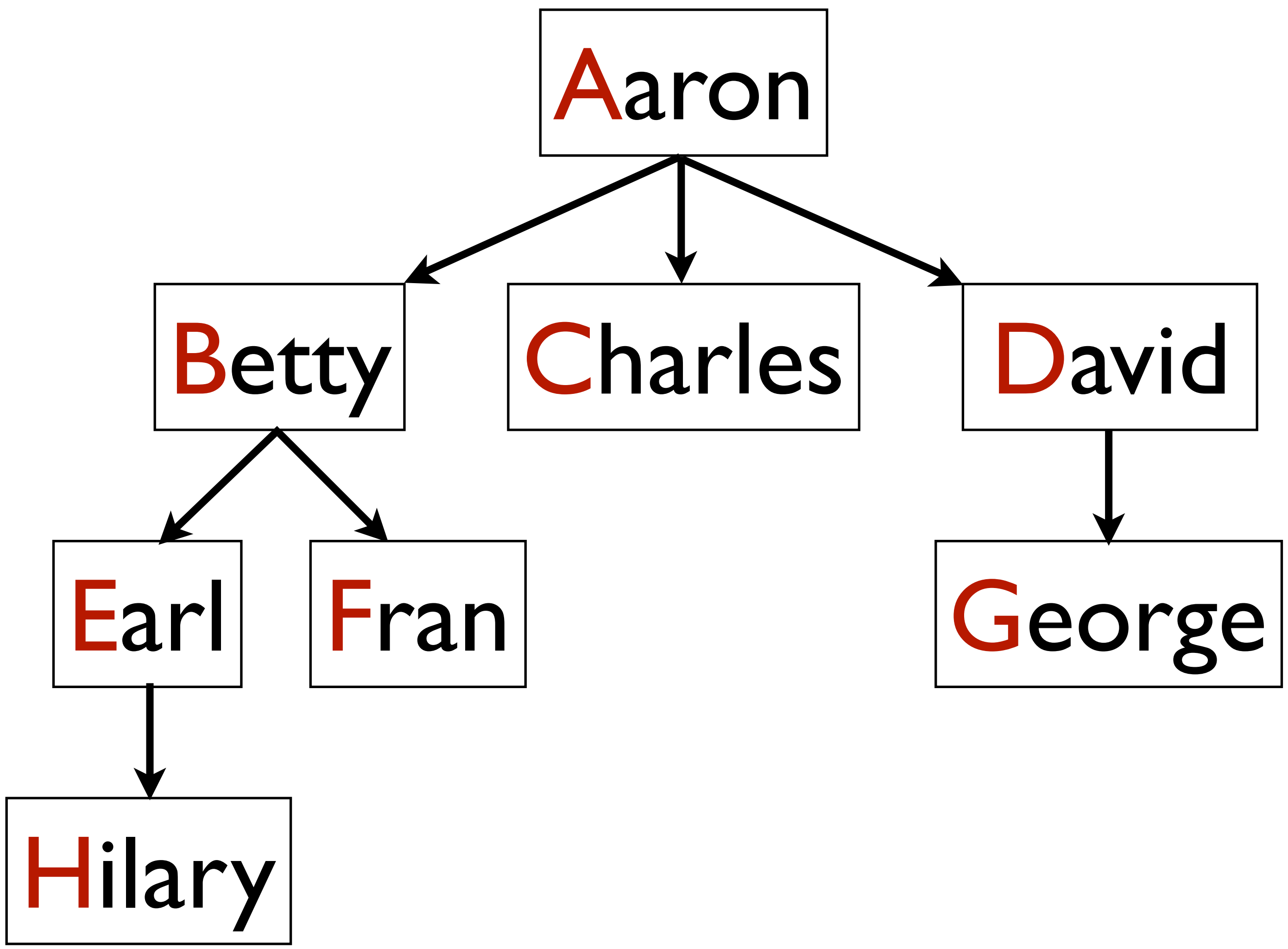
Tree-Revealing Process

Liben-Nowell, Kleinberg, PNAS'08



Tree-Revealing Process

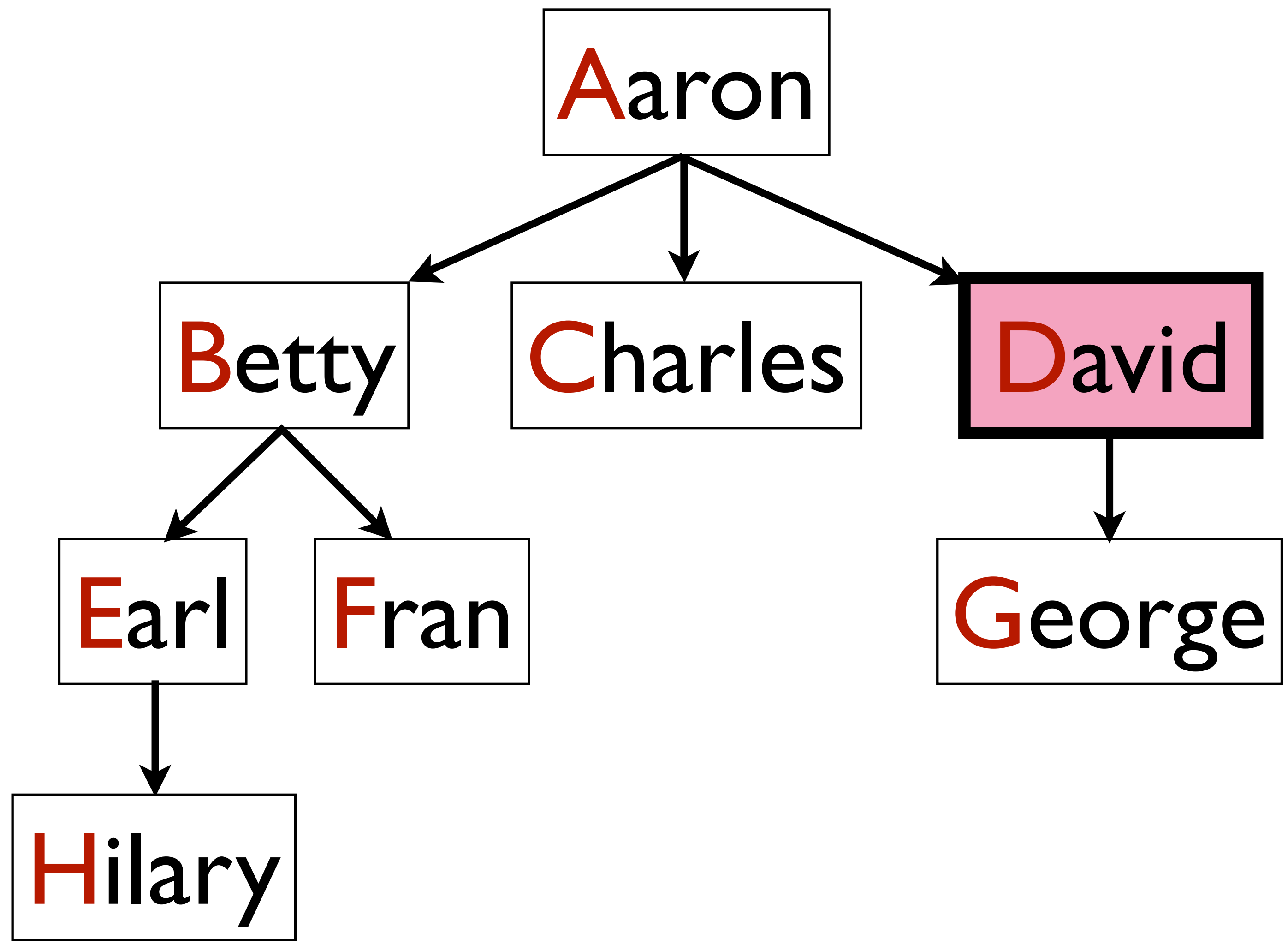
Liben-Nowell, Kleinberg, PNAS'08



Each node is exposed independently with prob. $\delta > 0$

Tree-Revealing Process

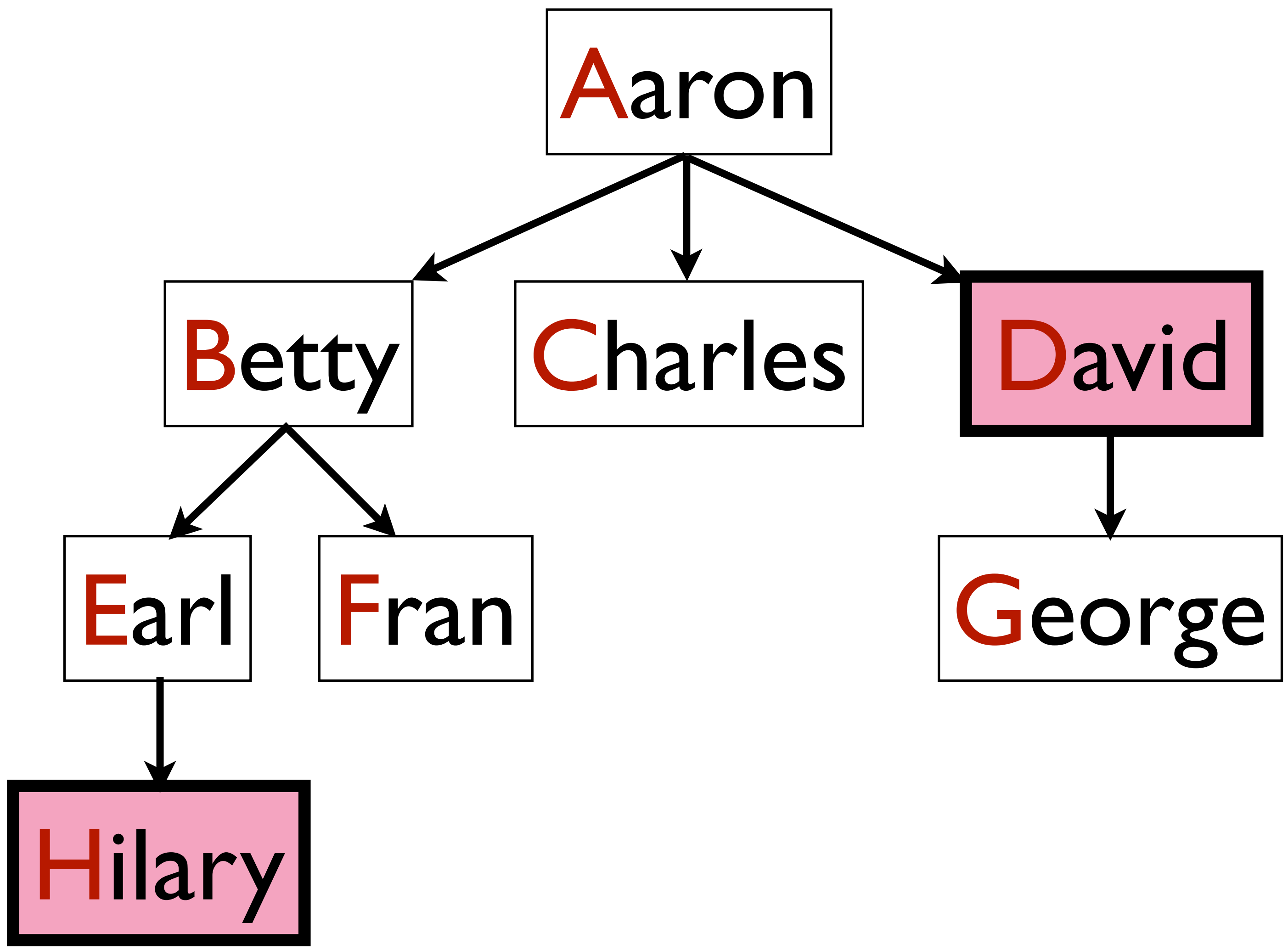
Liben-Nowell, Kleinberg, PNAS'08



Each node is exposed independently with prob. $\delta > 0$

Tree-Revealing Process

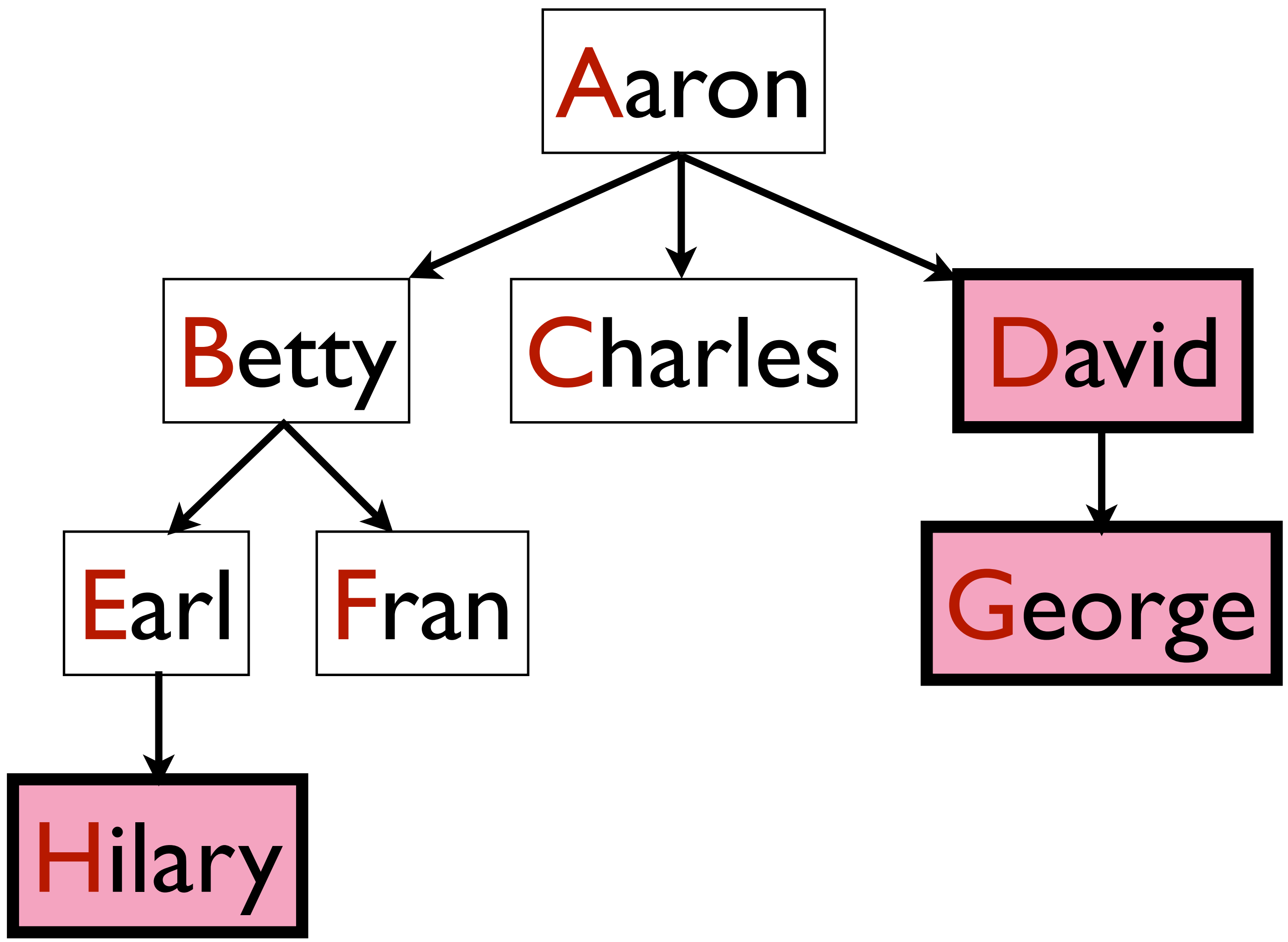
Liben-Nowell, Kleinberg, PNAS'08



Each node is exposed independently with prob. $\delta > 0$

Tree-Revealing Process

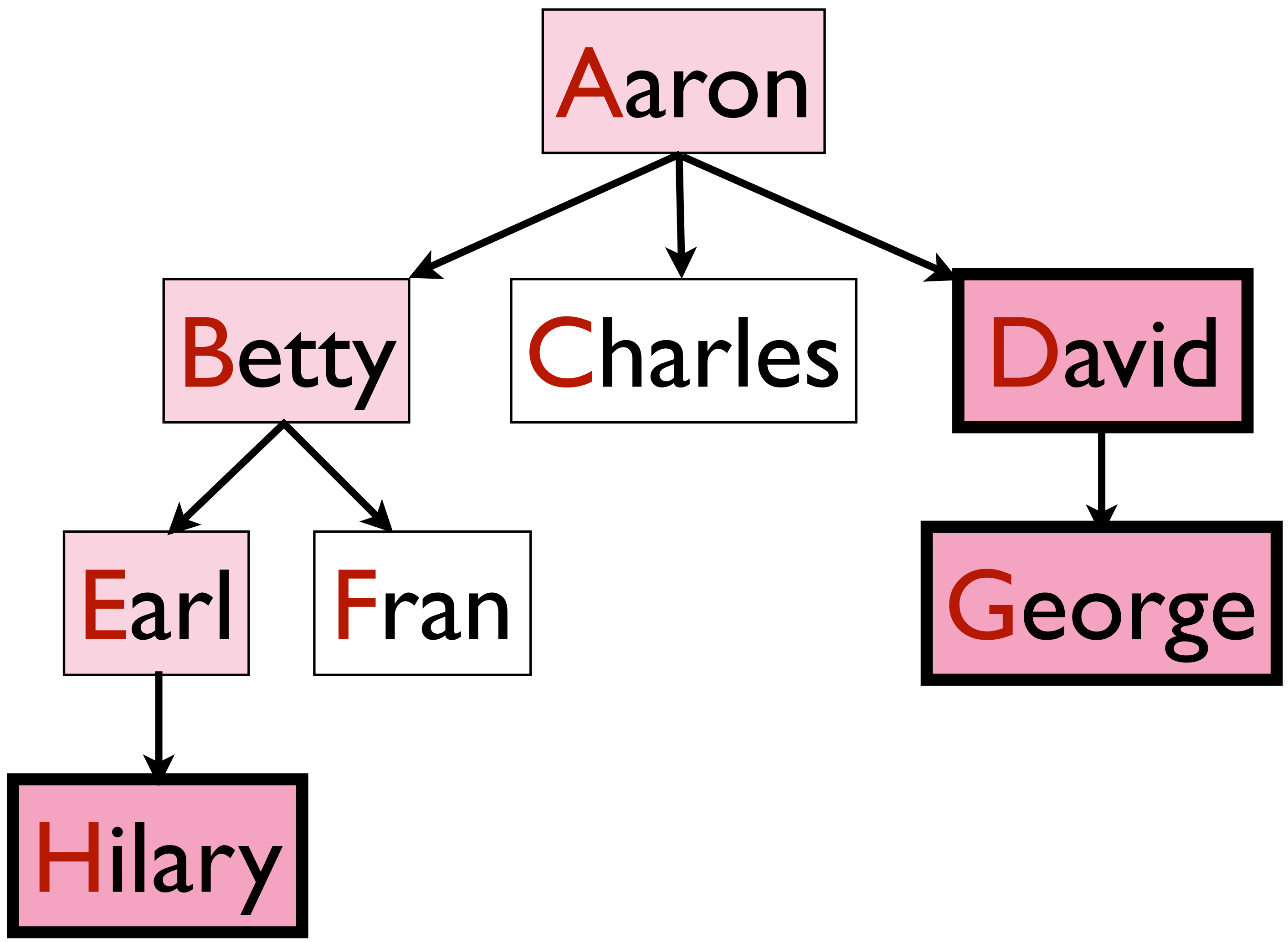
Liben-Nowell, Kleinberg, PNAS'08



Each node is exposed independently with prob. $\delta > 0$

Tree-Revealing Process

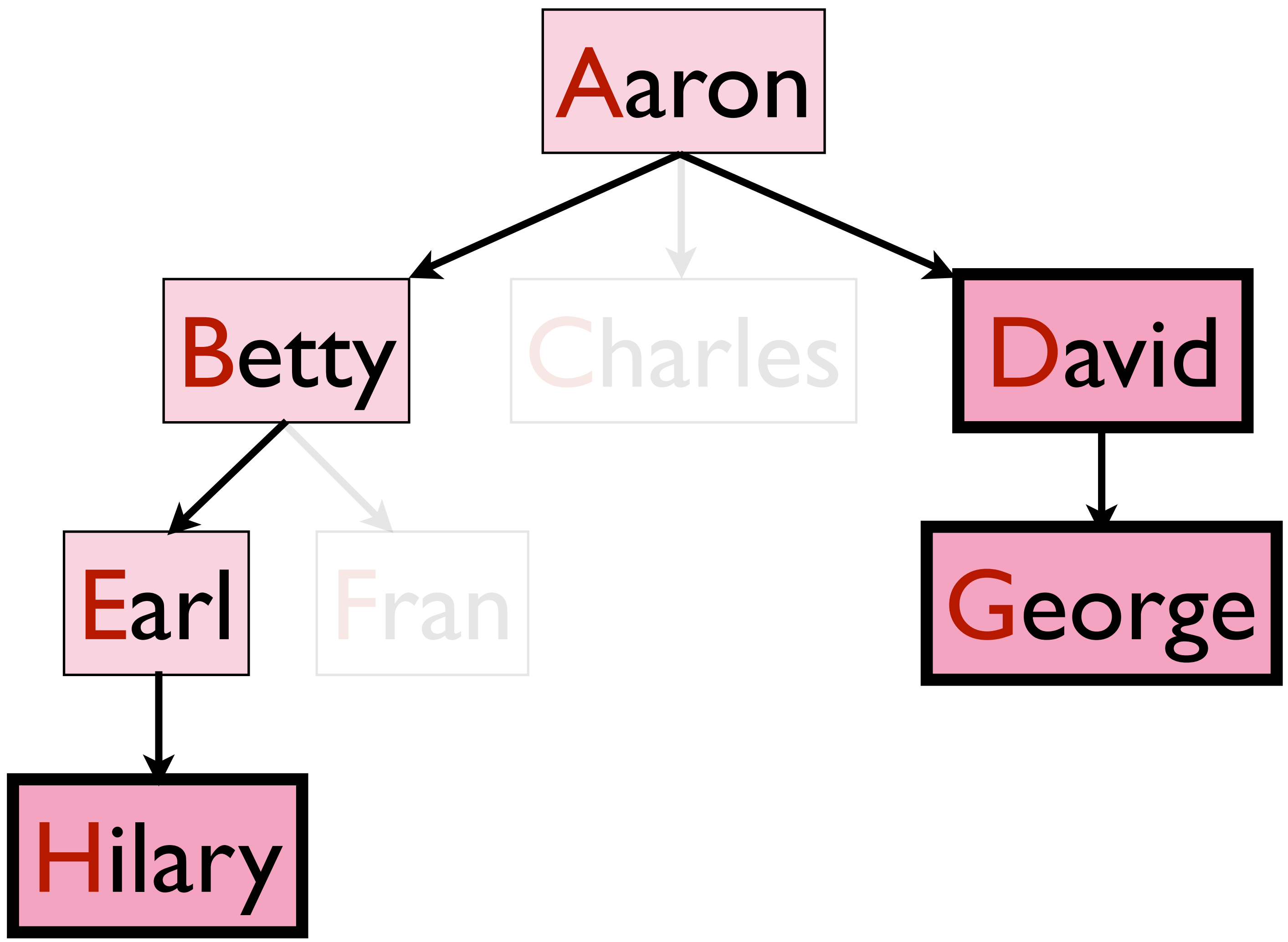
Liben-Nowell, Kleinberg, PNAS'08



Ancestors of exposed nodes are revealed

Tree-Revealing Process

Liben-Nowell, Kleinberg, PNAS'08



Ancestors of exposed nodes are revealed

Previous Work

- *Golub, Jackson, PNAS'10* perform simulations,
  - using *branching process trees* near the *critical threshold* as the Chain Letter Trees,
  - and *exposing* nodes as in *Kleinberg, Liben-Nowell, PNAS'08*.
- They observe that the *revealed* tree has a high fraction of nodes with only one child (and some other properties).

Our Contribution

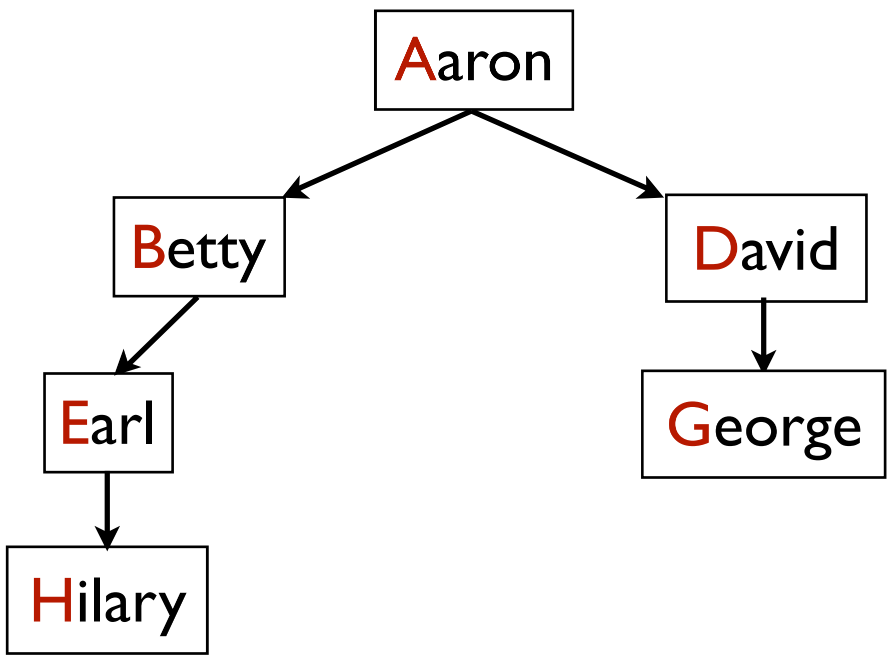
- Our 1st result, *informally*, states that the tree-revealing process, is enough to explain the high fraction of single-child nodes

Our Contribution

- Our 1st result, *informally*, states that the tree-revealing process, is enough to explain the high fraction of single-child nodes assuming only a degree bound on the unknown chain letter tree.

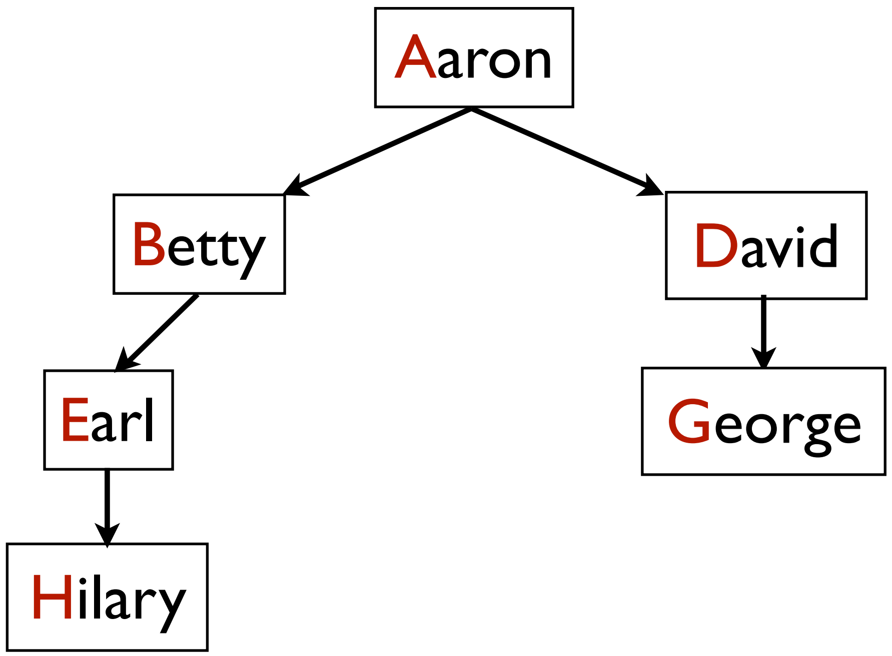
Revealed vs. Unknown

We see a
“revealed” tree...

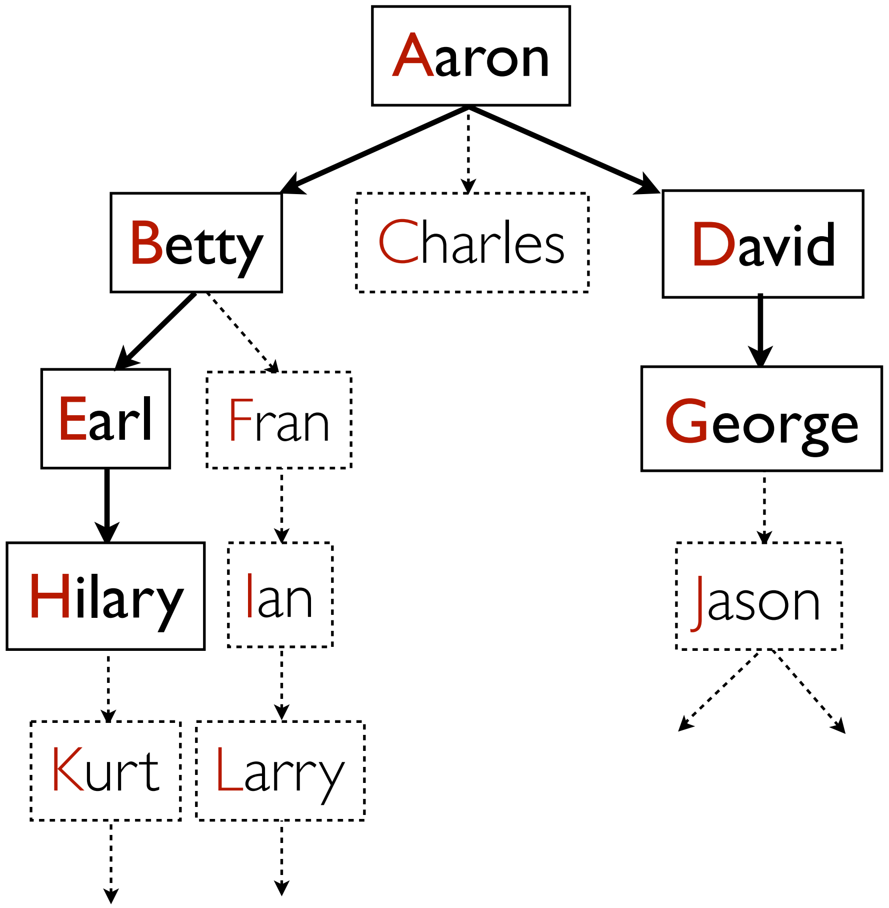


Revealed vs. Unknown

We see a "revealed" tree...

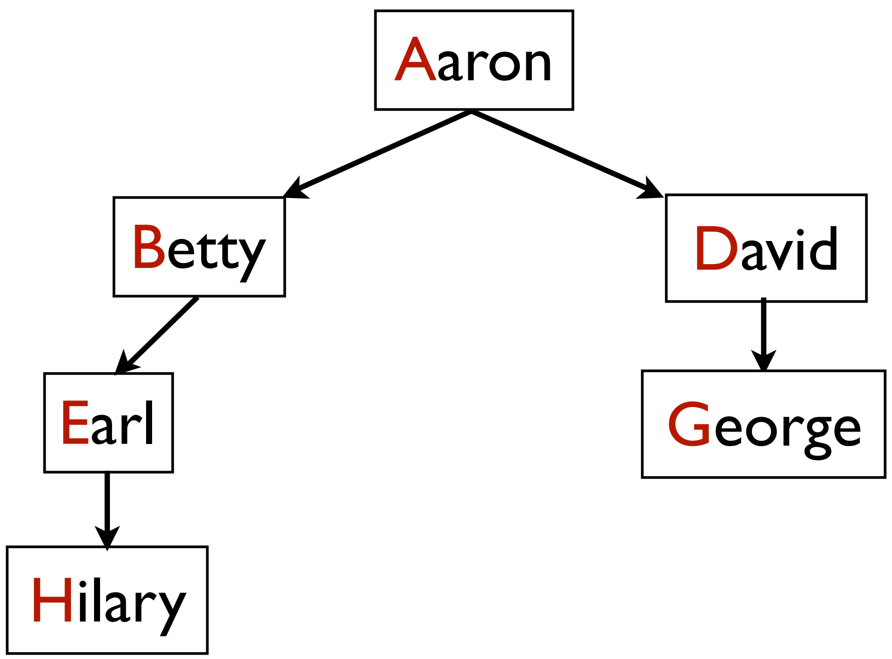


...we would like to study the "unknown" tree!

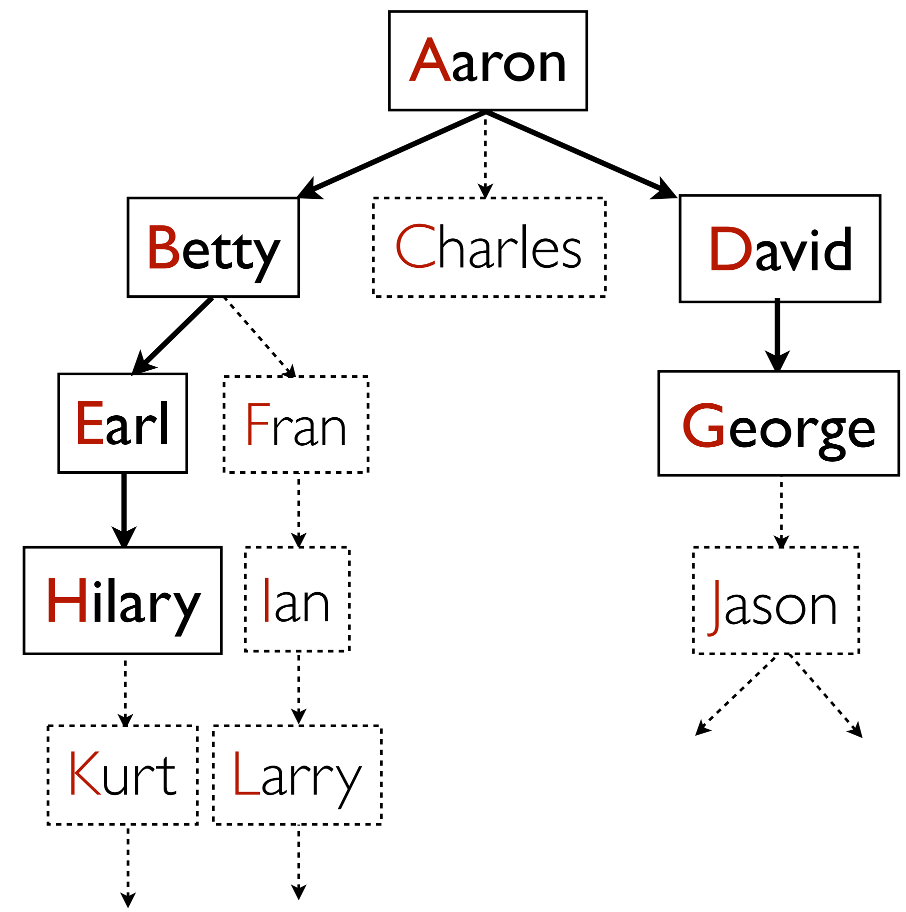


Revealed vs. Unknown

We see a "revealed" tree...



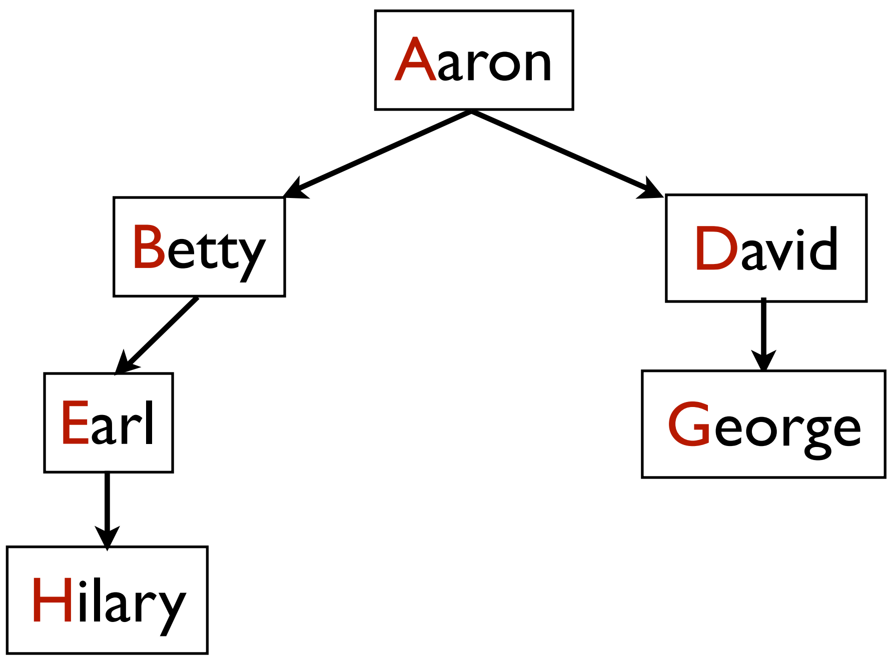
...we would like to study the "unknown" tree!



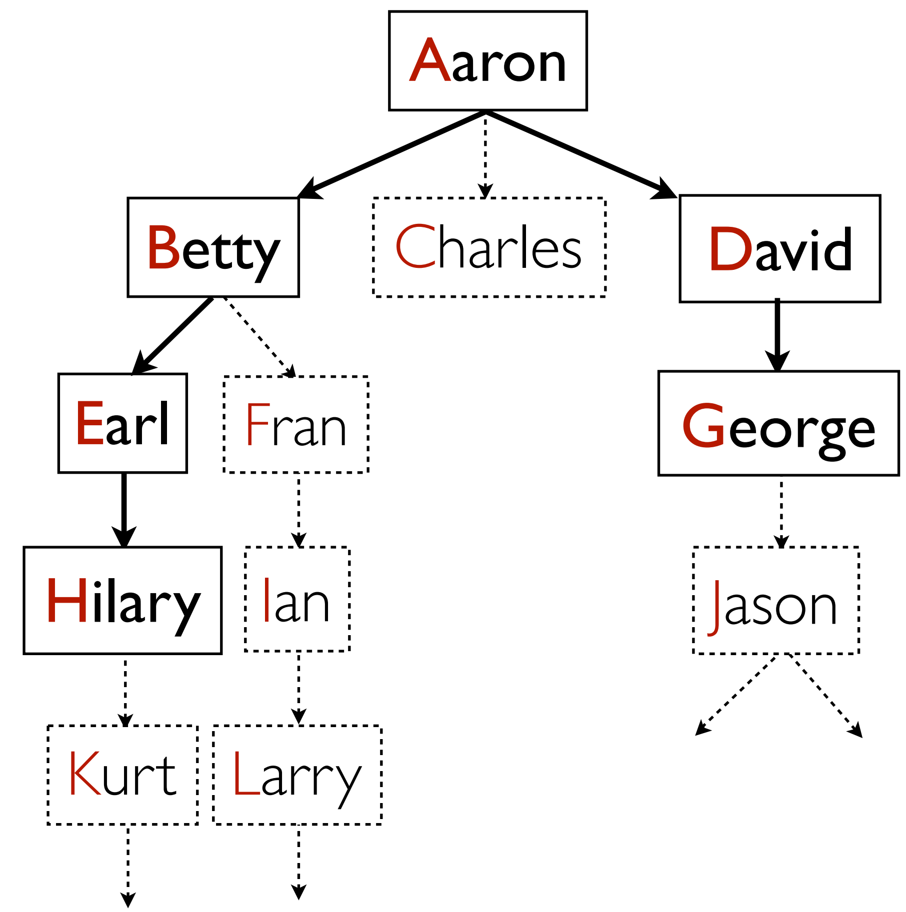
Size? Width? Height? Degree Distribution? ...

Revealed vs. Unknown

We see a "revealed" tree...



...we would like to study the "unknown" tree!



$\boxed{\text{Size?}}$ Width? Height? Degree Distribution? ...

Our Contribution

- Our 2nd result, *informally*, states that (under reasonable assumptions) it is possible to estimate the size of the unknown chain letter tree with a small error, with high probability.

Our Contribution

- Our 2nd result, *informally*, states that (under reasonable assumptions) it is possible to estimate the size of the unknown chain letter tree with a small error, with high probability.

Observe that we do not know the exposing probability δ

Our Contribution

- Our 2nd result, *informally*, states that (under reasonable assumptions) it is possible to estimate the size of the unknown chain letter tree with a small error, with high probability.

We use this theorem to estimate that ~ 173k people that signed the IRAQ chain letter

This estimate is backed by a probability bound (on the probability space induced by the revealing process)

Our Contribution

- Our 2nd result, *informally*, states that (under reasonable assumptions) it is possible to estimate the size of the unknown chain letter tree with a small error, with high probability.

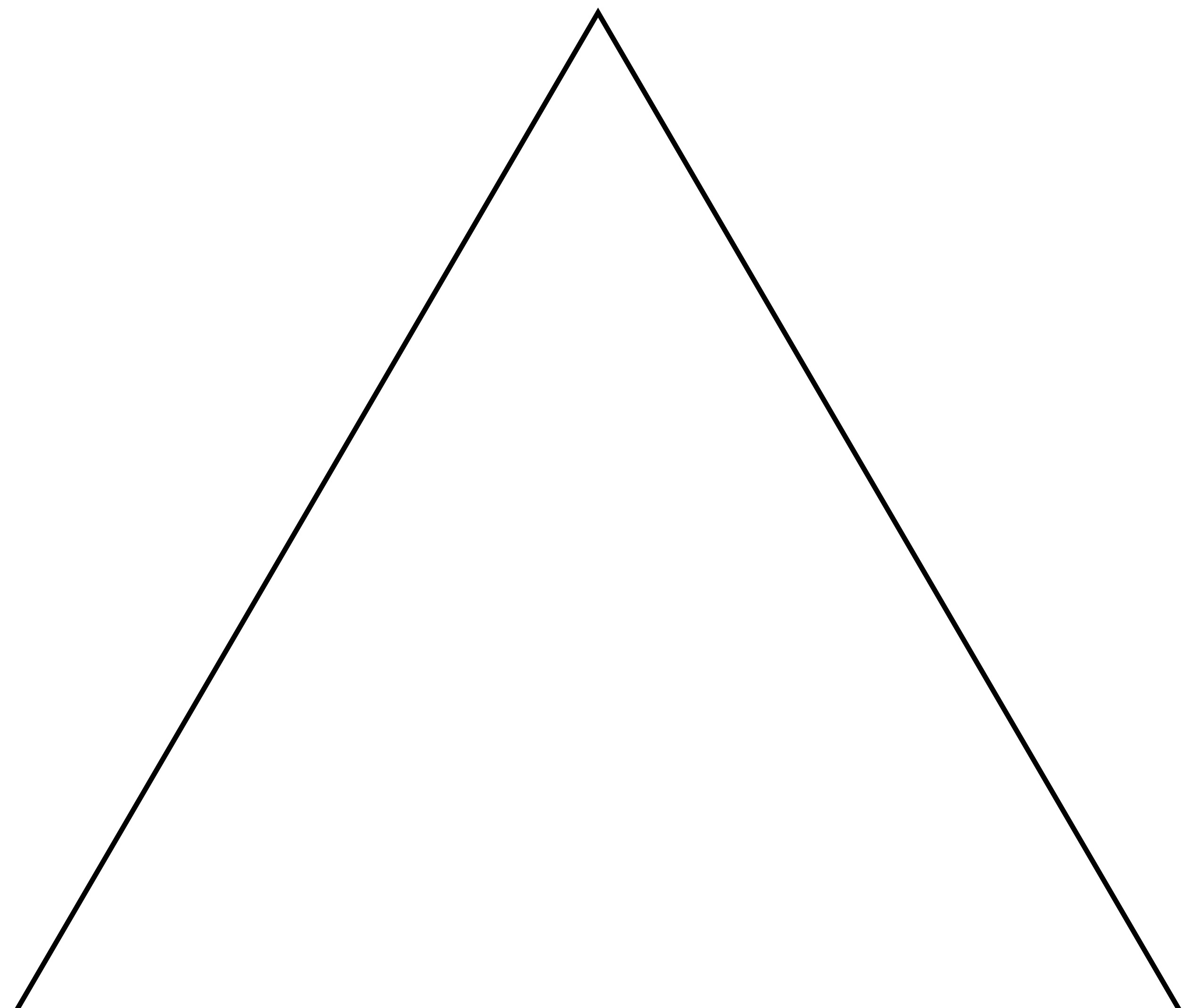
We use this theorem to estimate that ~ 173k people that signed the IRAQ chain letter

The chain letter generated ~ 3.5M emails

Single-Child Fraction

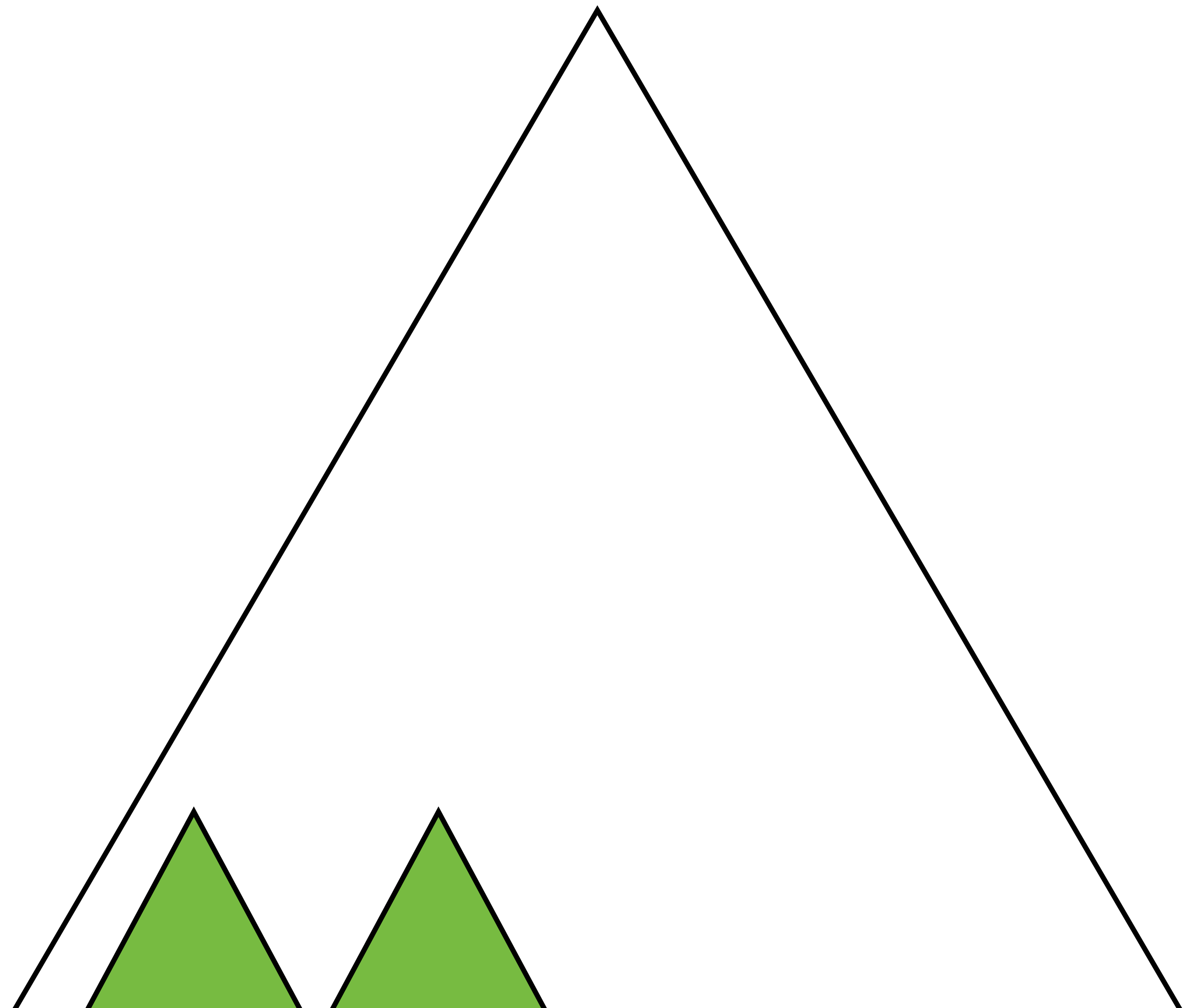
- Nodes are exposed with probability $\delta > 0$
- We assume that the unknown tree's maximum degree is at most k

Single-Child Fraction



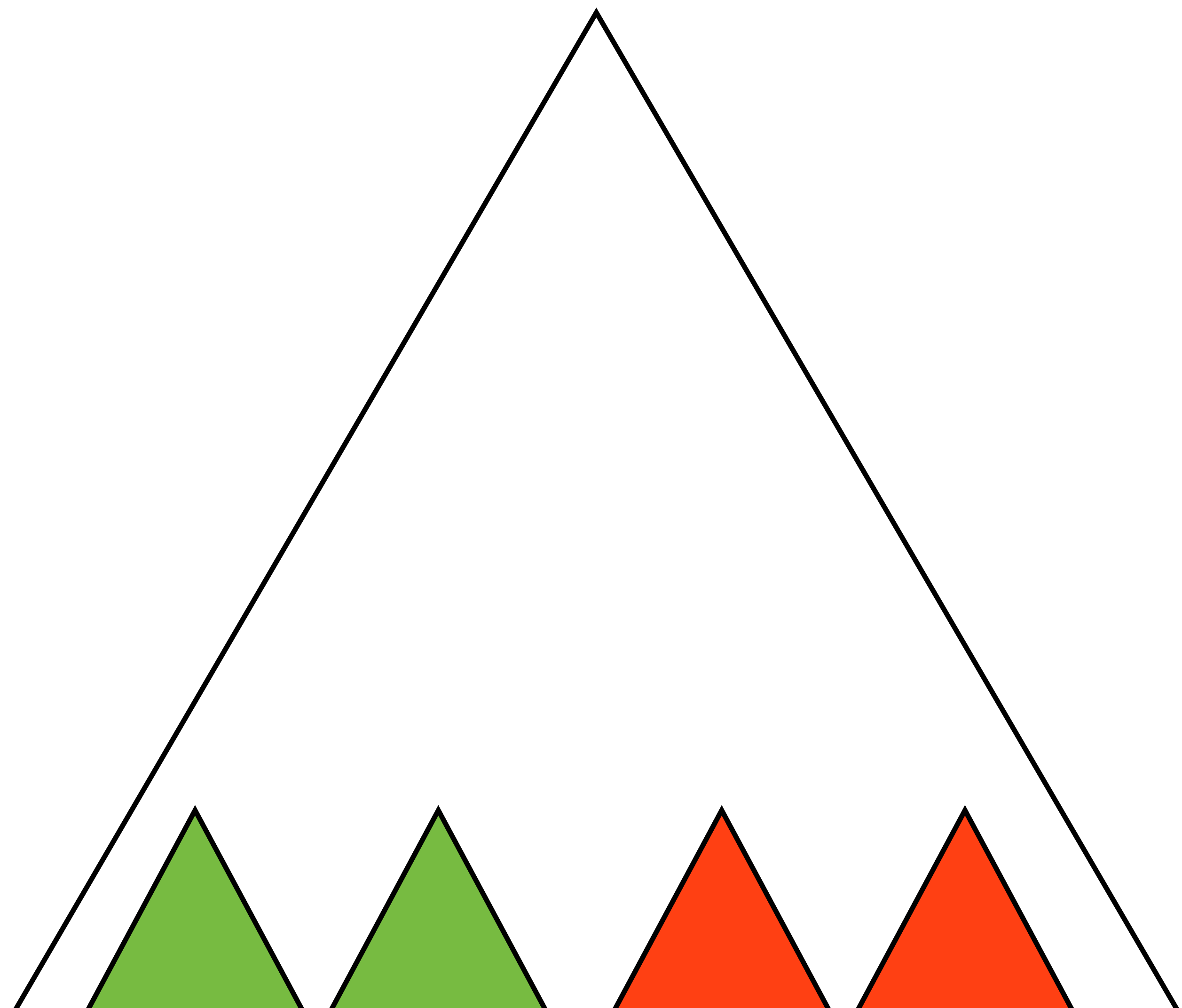
We partition the tree into subforests

Single-Child Fraction



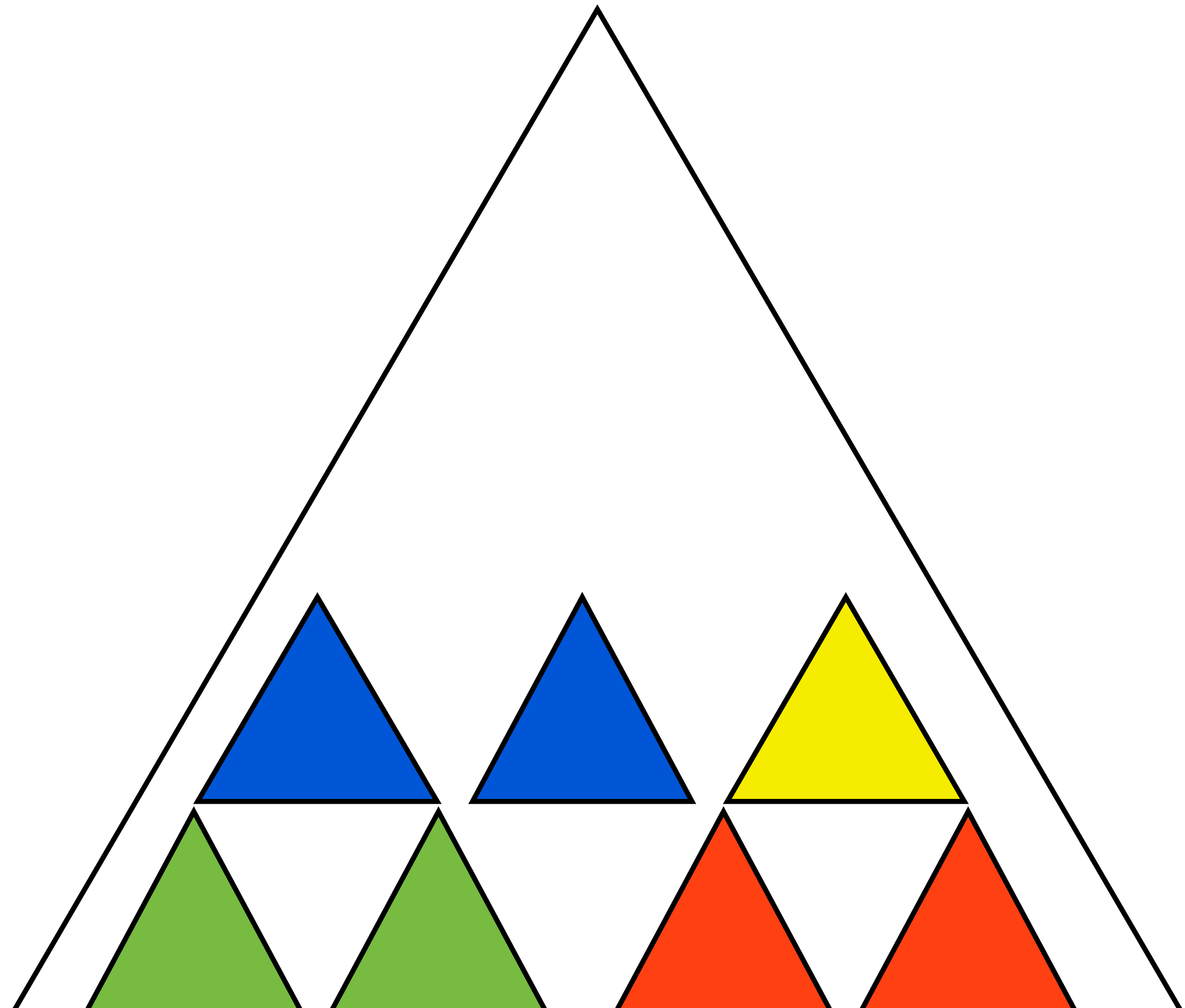
We partition the tree into subforests

Single-Child Fraction



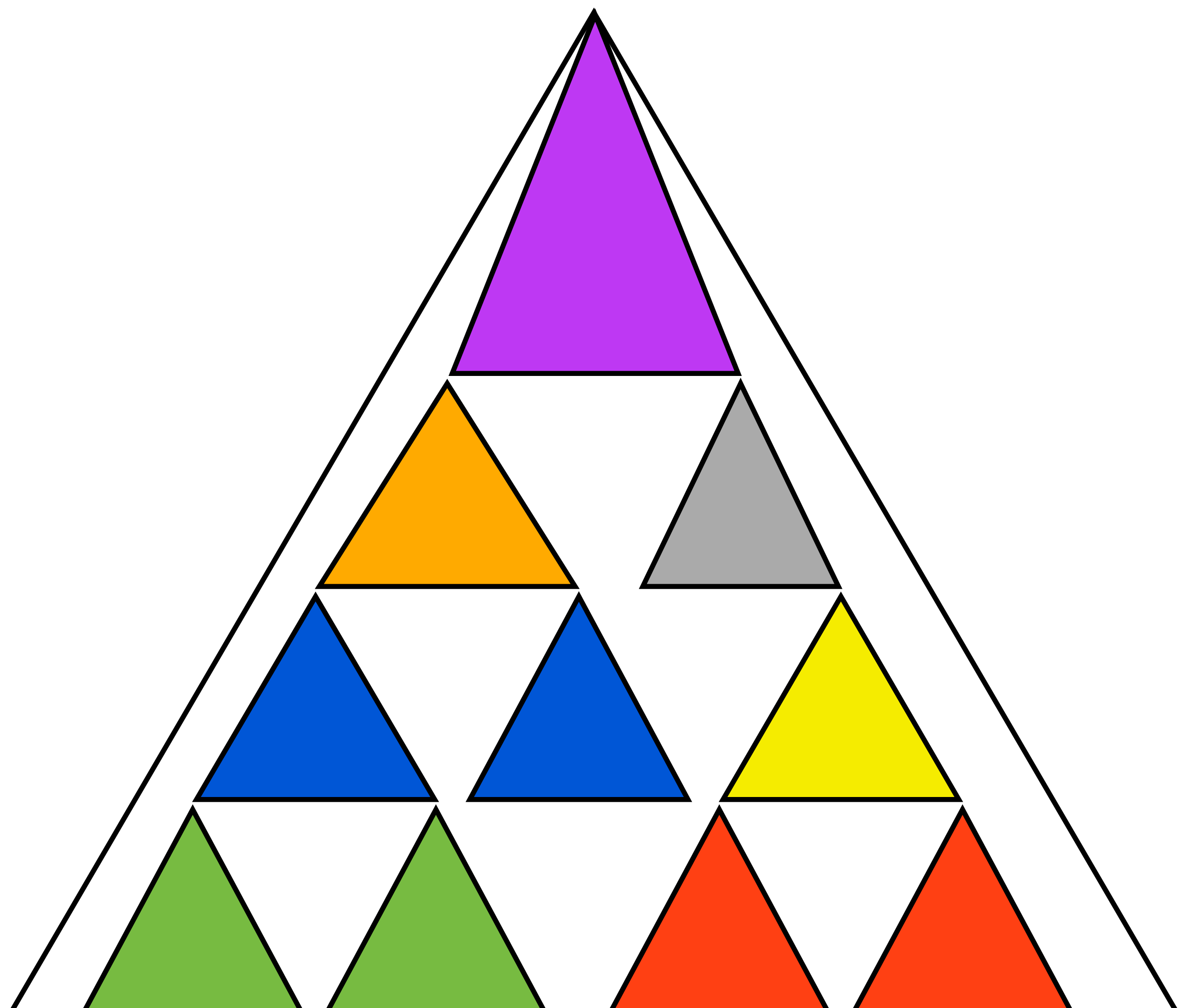
We partition the tree into subforests

Single-Child Fraction



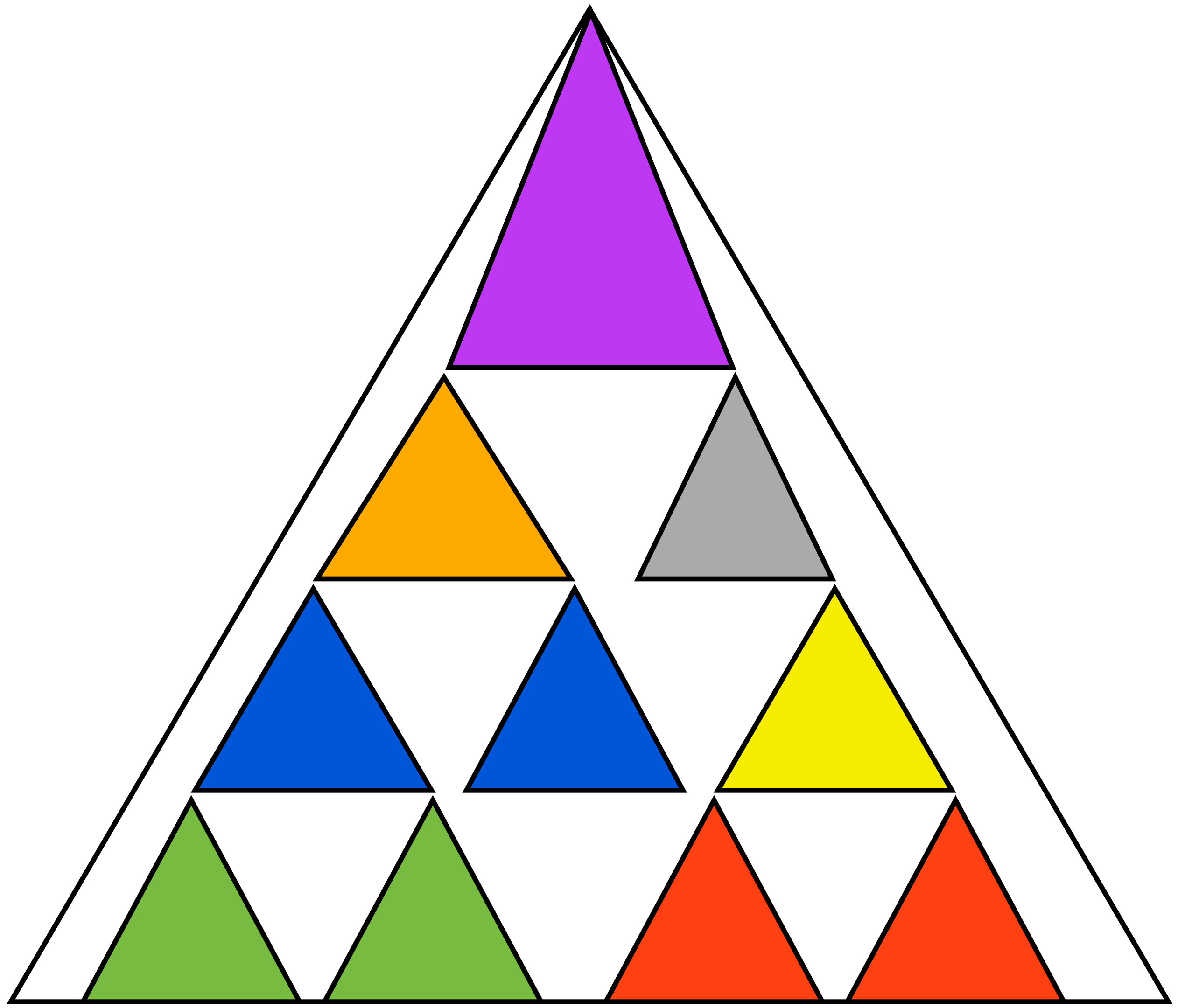
We partition the tree into subforests

Single-Child Fraction



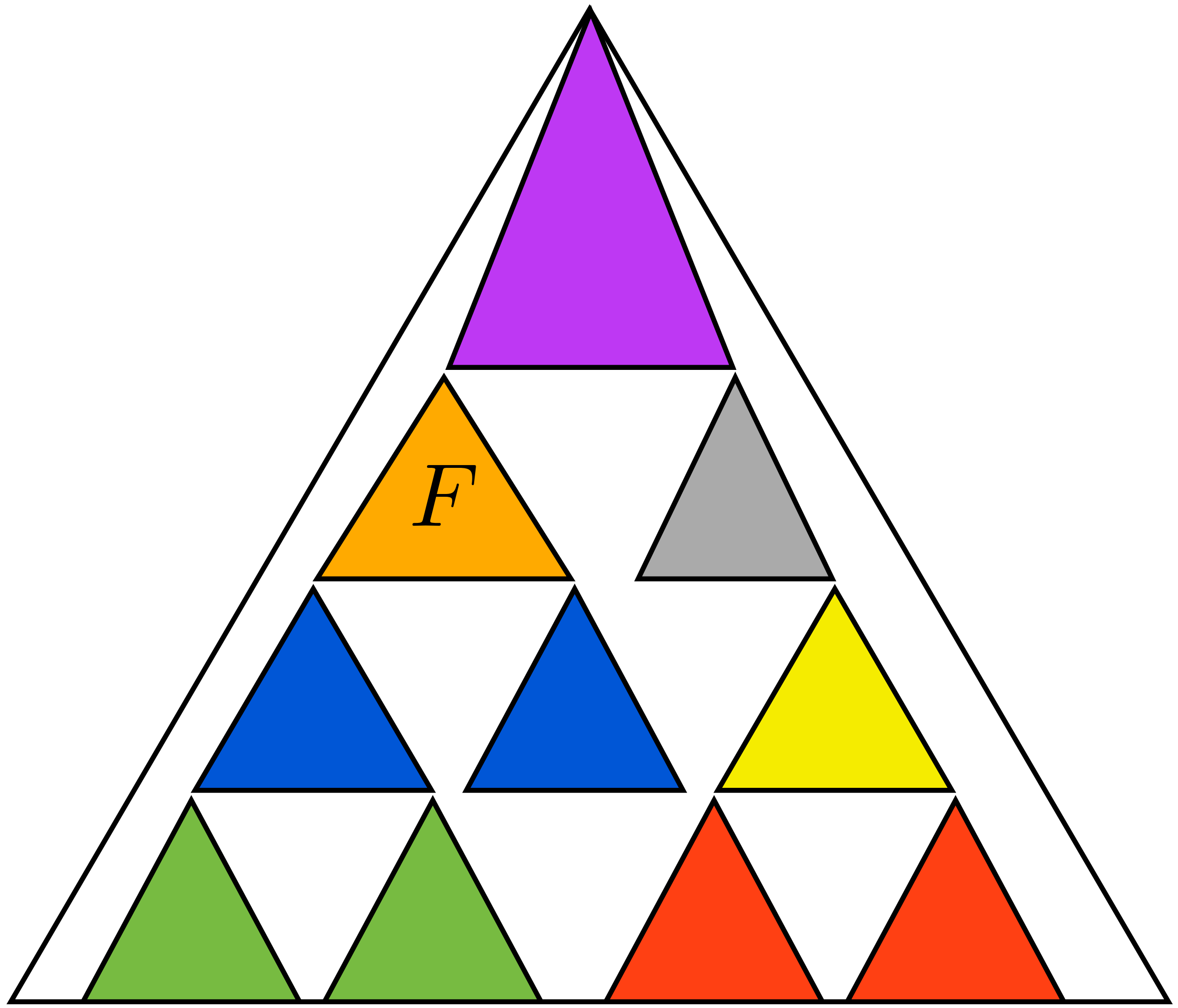
We partition the tree into subforests

Single-Child Fraction



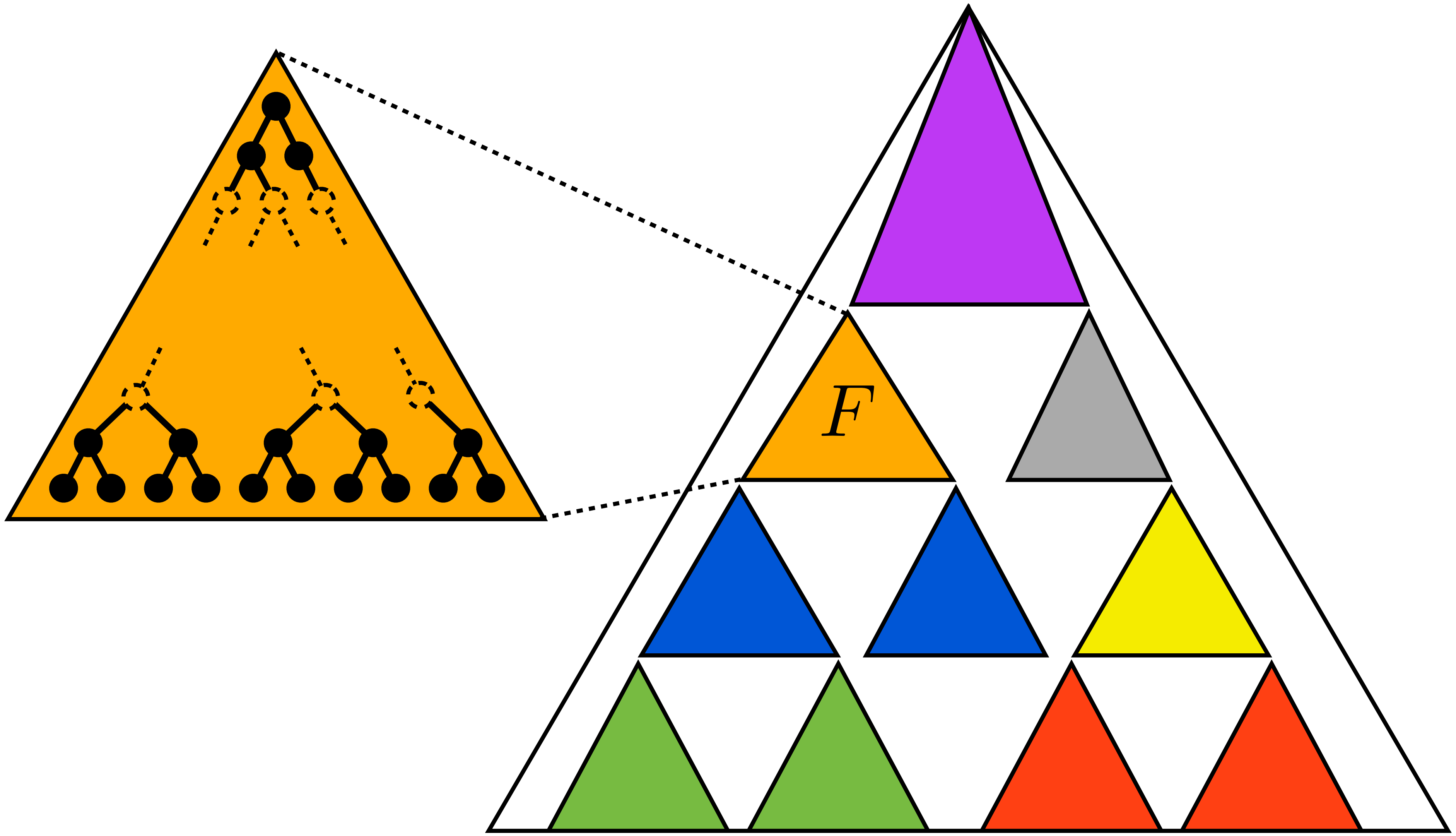
We partition the tree into subforests,
in such a way that each subforest has $\simeq \delta^{-1}$ nodes
and the median height in the subforest is $\Omega\left(\log_{k-1} \delta^{-1}\right)$.

Single-Child Fraction



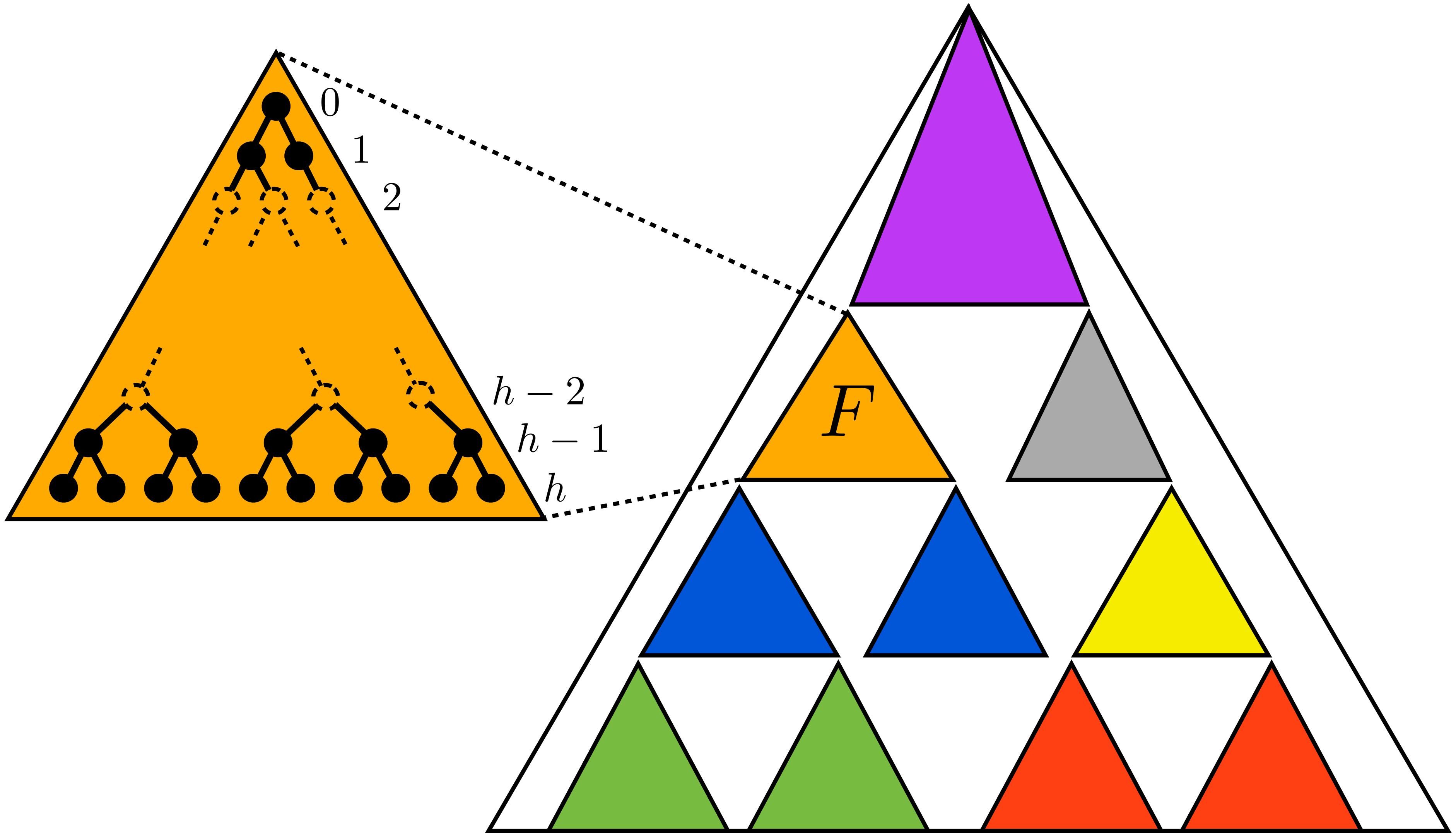
We partition the tree into subforests,
in such a way that each subforest has $\simeq \delta^{-1}$ nodes
and the median height in the subforest is $\Omega\left(\log_{k-1} \delta^{-1}\right)$.

Single-Child Fraction



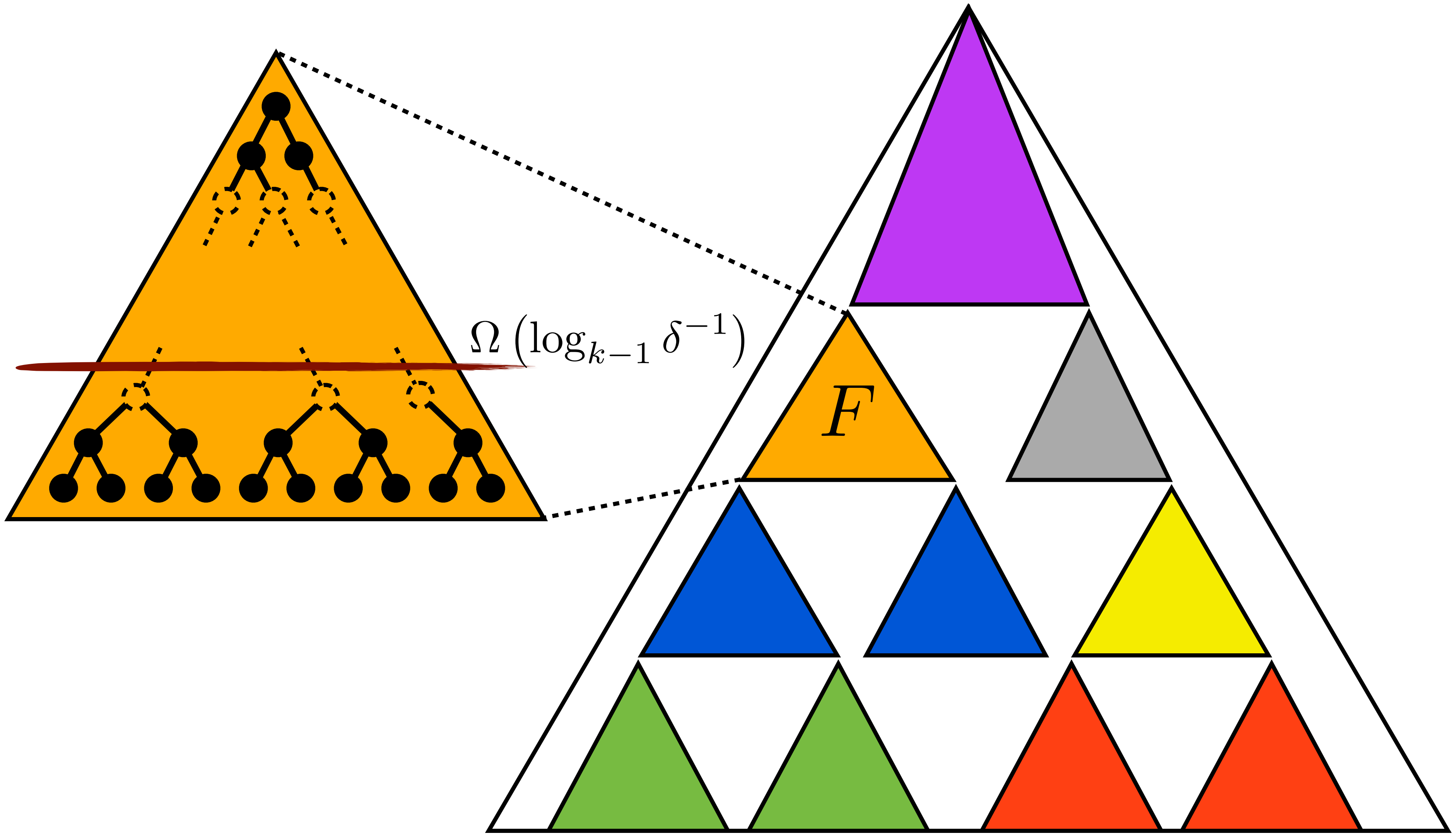
We partition the tree into subforests,
in such a way that each subforest has $\simeq \delta^{-1}$ nodes
and the median height in the subforest is $\Omega\left(\log_{k-1} \delta^{-1}\right)$.

Single-Child Fraction



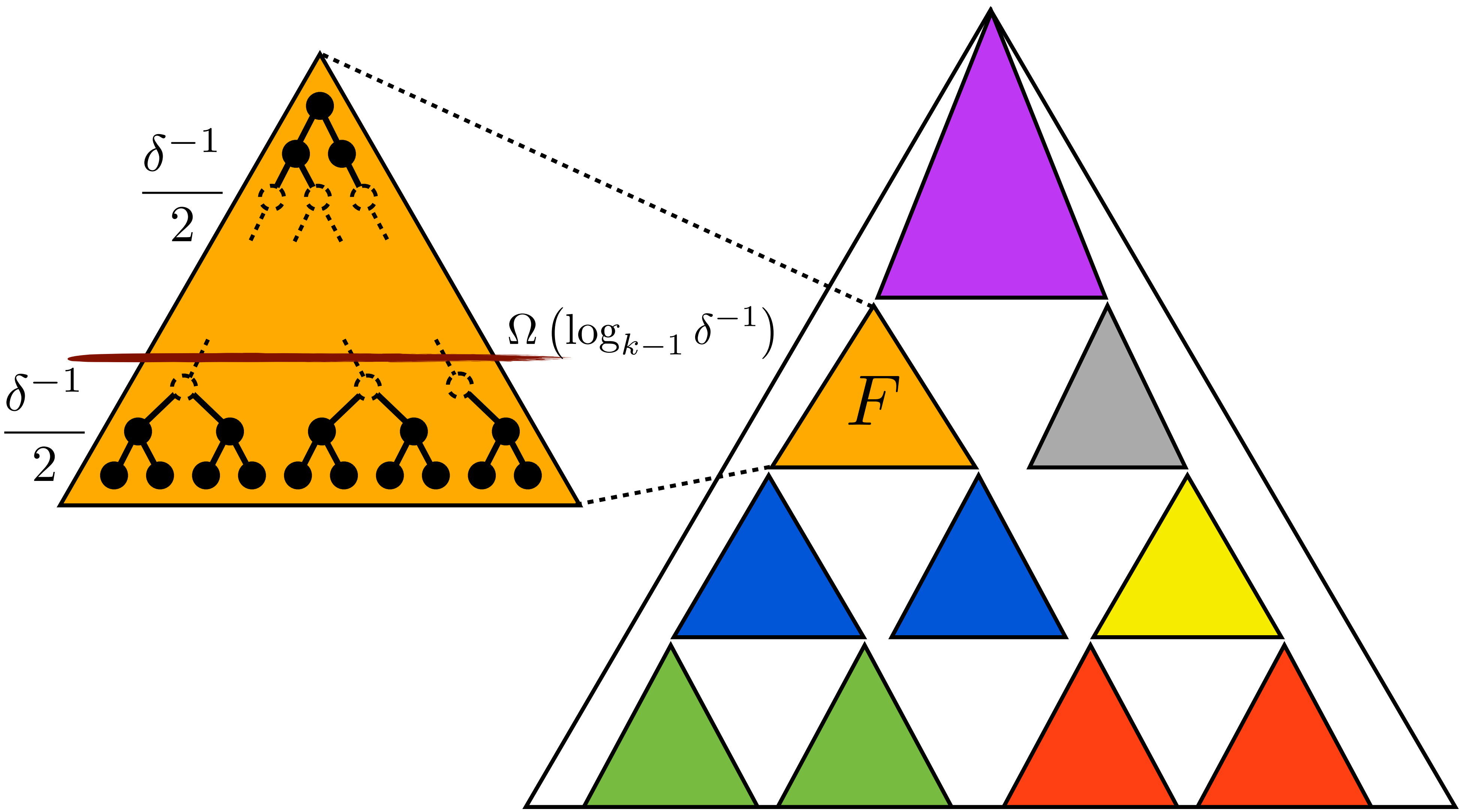
We partition the tree into subforests,
in such a way that each subforest has $\simeq \delta^{-1}$ nodes
and the median height in the subforest is $\Omega\left(\log_{k-1} \delta^{-1}\right)$.

Single-Child Fraction



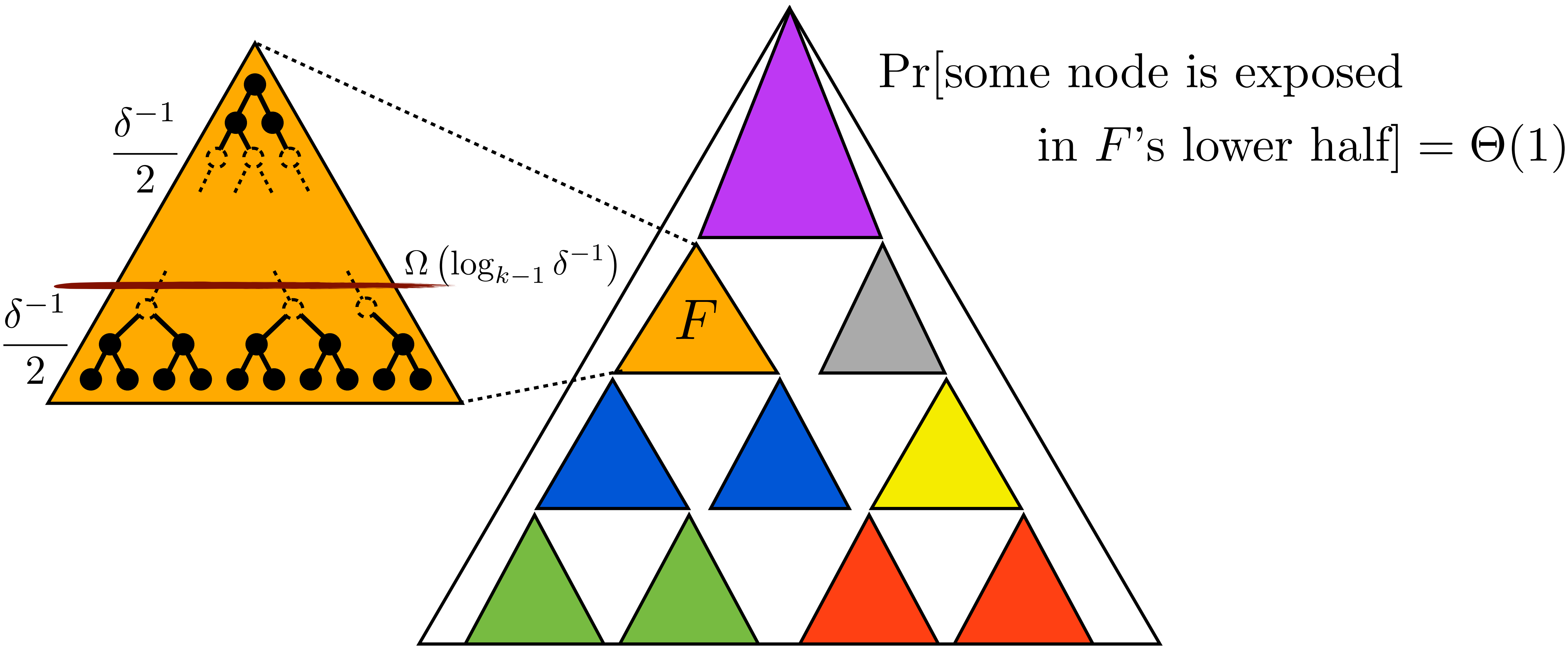
We partition the tree into subforests,
in such a way that each subforest has $\simeq \delta^{-1}$ nodes
and the median height in the subforest is $\Omega\left(\log_{k-1} \delta^{-1}\right)$.

Single-Child Fraction



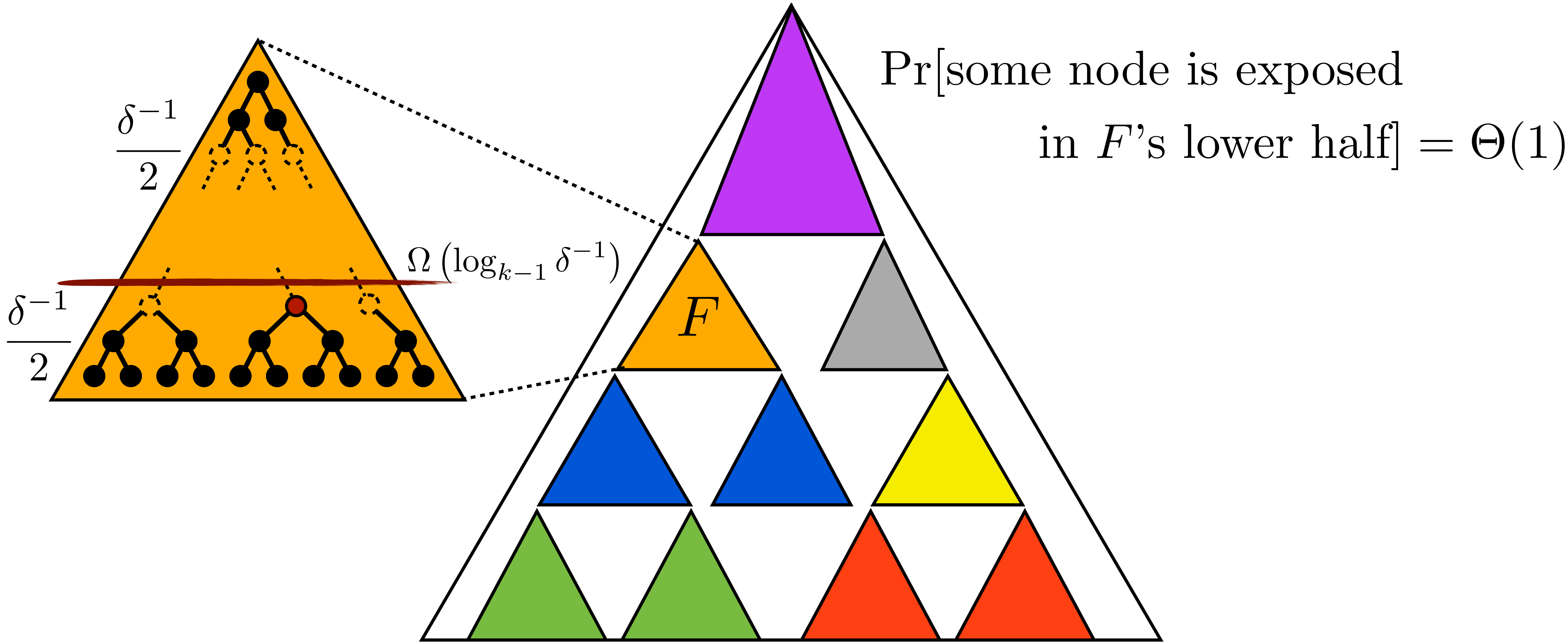
We partition the tree into subforests,
in such a way that each subforest has $\simeq \delta^{-1}$ nodes
and the median height in the subforest is $\Omega\left(\log_{k-1}\delta^{-1}\right)$.

Single-Child Fraction



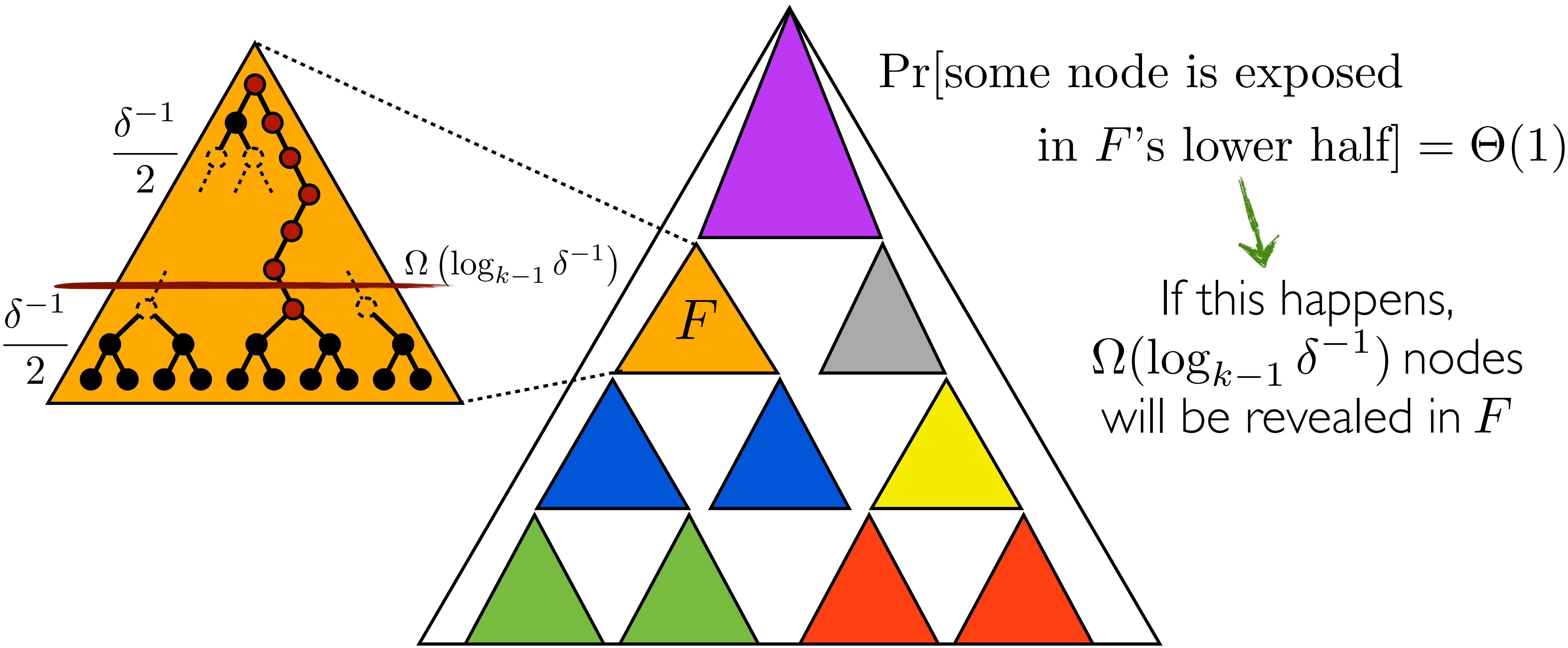
We partition the tree into subforests, in such a way that each subforest has $\simeq \delta^{-1}$ nodes and the median height in the subforest is $\Omega\left(\log_{k-1}\delta^{-1}\right)$.

Single-Child Fraction



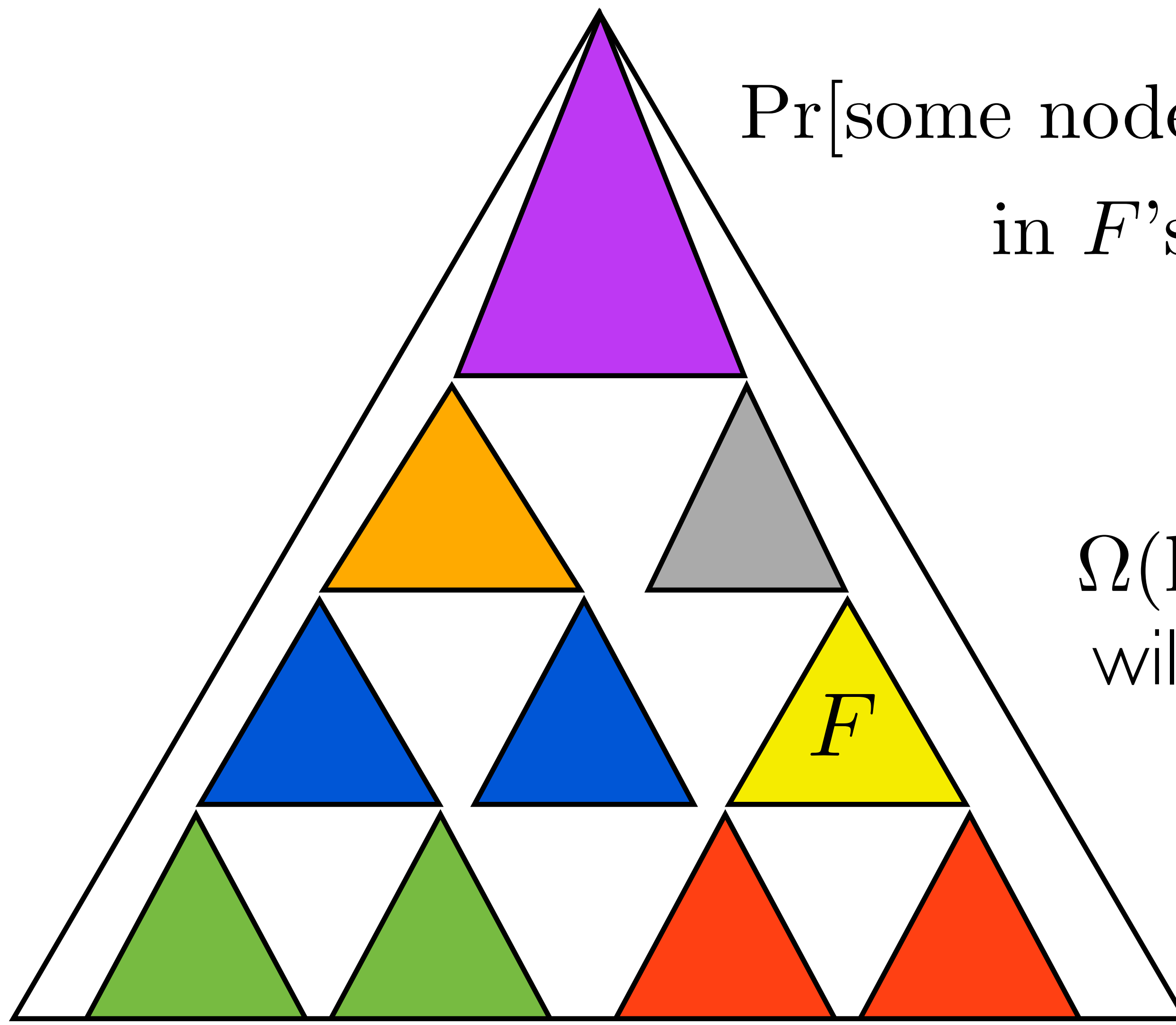
We partition the tree into subforests,
in such a way that each subforest has $\simeq \delta^{-1}$ nodes
and the median height in the subforest is $\Omega\left(\log_{k-1}\delta^{-1}\right)$.

Single-Child Fraction



We partition the tree into subforests, in such a way that each subforest has $\simeq \delta^{-1}$ nodes and the median height in the subforest is $\Omega\left(\log_{k-1}\delta^{-1}\right)$.

Single-Child Fraction

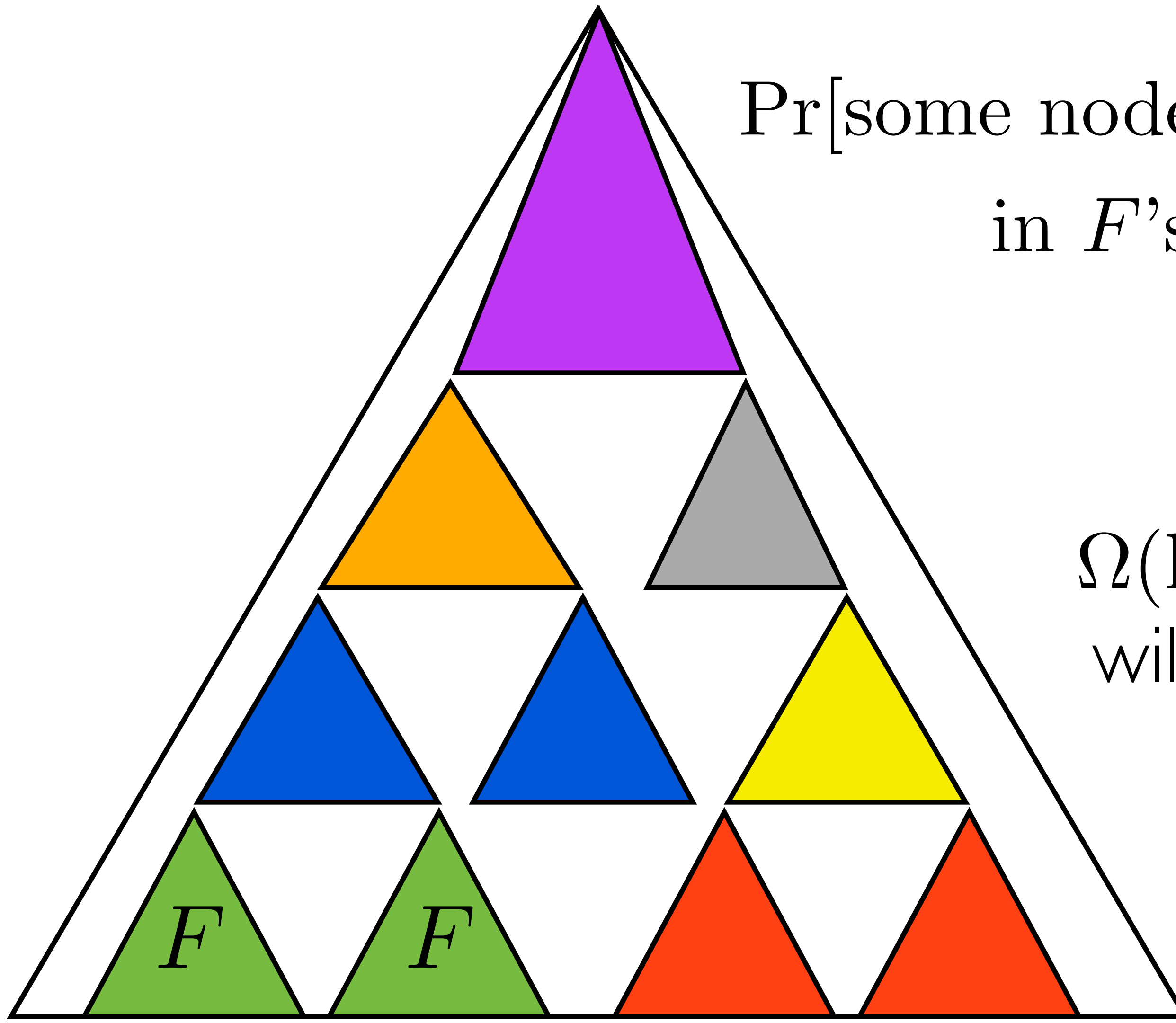


$\Pr[\text{some node is exposed in } F\text{'s lower half}] = \Theta(1)$

If this happens,
$\Omega(\log_{k-1} \delta^{-1})$ nodes
will be revealed in F

We partition the tree into subforests,
in such a way that each subforest has $\simeq \delta^{-1}$ nodes
and the median height in the subforest is $\Omega\left(\log_{k-1} \delta^{-1}\right)$.

Single-Child Fraction



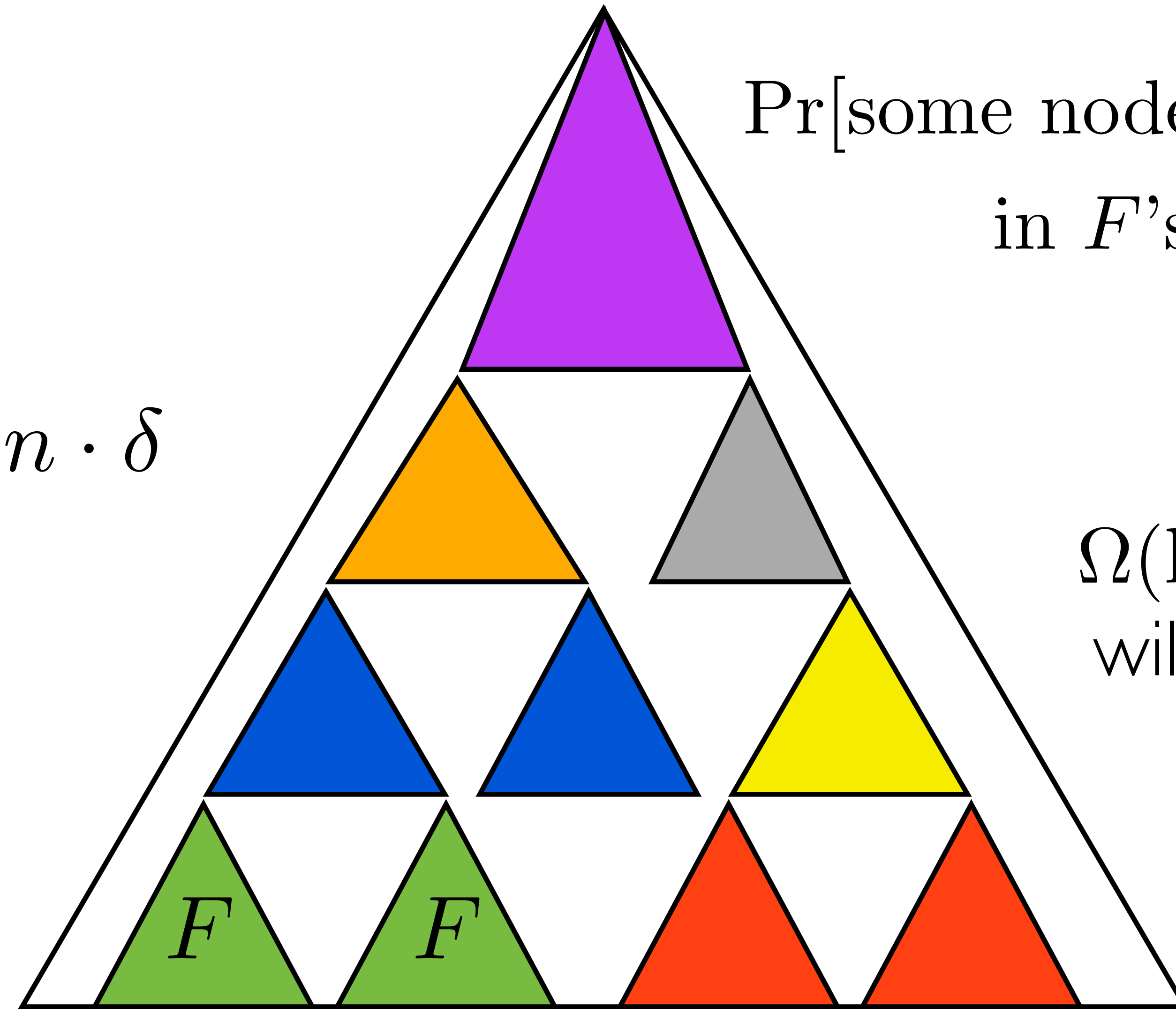
$\Pr[$some node is exposed in F's lower half$] = \Theta(1)$

If this happens, $\Omega(\log_{k-1} \delta^{-1})$ nodes will be revealed in F

We partition the tree into subforests, in such a way that each subforest has $\simeq \delta^{-1}$ nodes and the median height in the subforest is $\Omega\left(\log_{k-1} \delta^{-1}\right)$.

Single-Child Fraction

$\#$ of forests $\simeq n \cdot \delta$



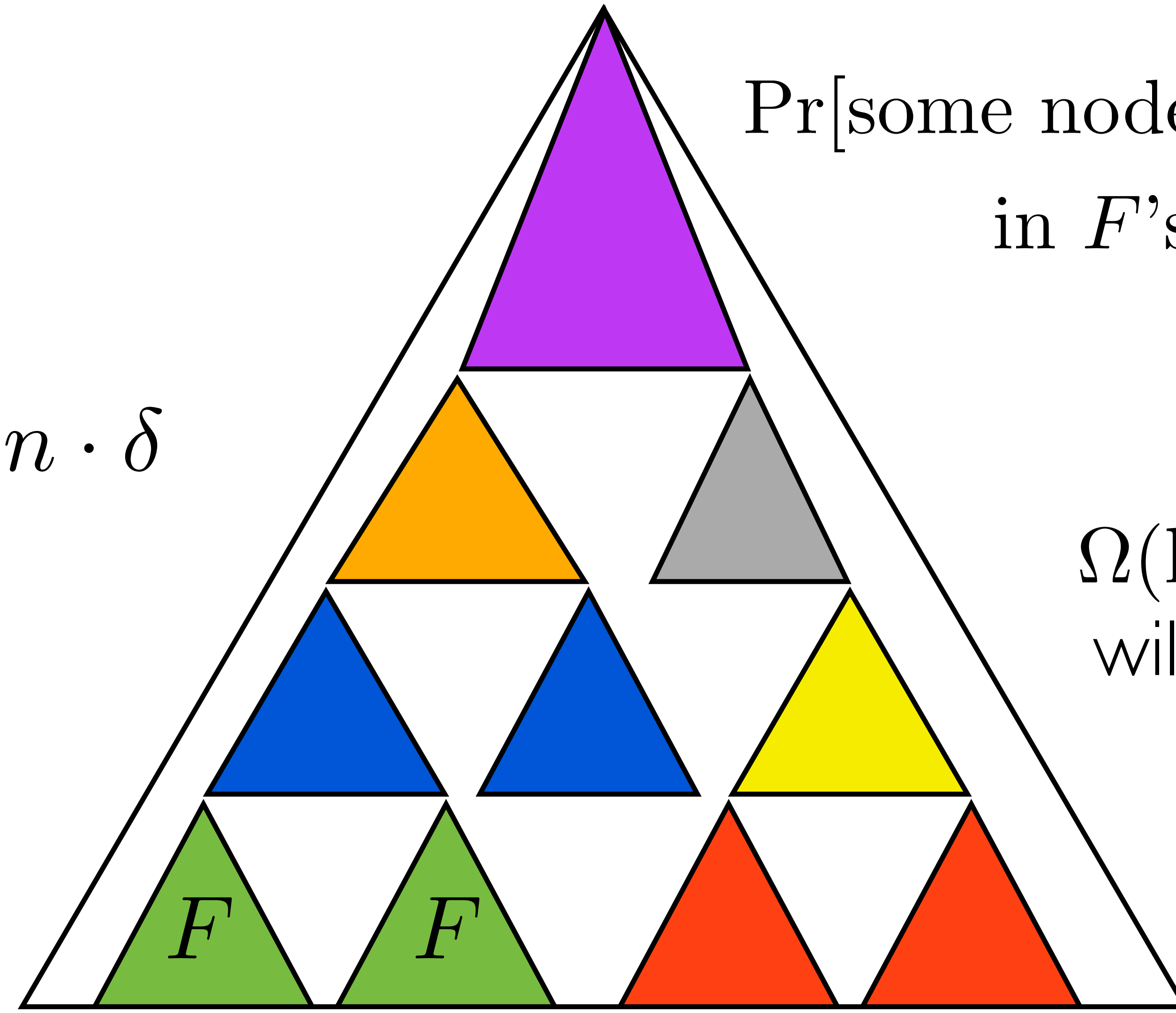
$\Pr[$some node is exposed in F's lower half$] = \Theta(1)$

If this happens, $\Omega(\log_{k-1} \delta^{-1})$ nodes will be revealed in F

We partition the tree into subforests, in such a way that each subforest has $\simeq \delta^{-1}$ nodes and the median height in the subforest is $\Omega\left(\log_{k-1} \delta^{-1}\right)$.

Single-Child Fraction

$\#$ of forests $\simeq n \cdot \delta$



Pr[some node is exposed in F's lower half] $= \Theta(1)$

If this happens, $\Omega(\log_{k-1} \delta^{-1})$ nodes will be revealed in F

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

Single-Child Fraction

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

Single-Child Fraction

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

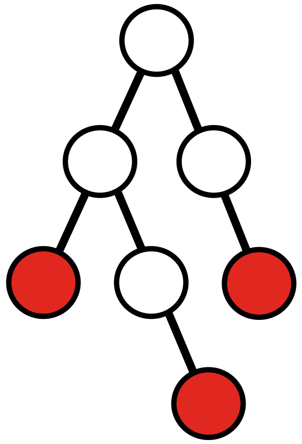
$$\Pr[\text{at most } 2 \cdot n \cdot \delta \text{ nodes will be exposed}] = 1 - o(1)$$

Single-Child Fraction

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

$$\Pr[\text{at most } 2 \cdot n \cdot \delta \text{ nodes will be exposed}] = 1 - o(1)$$

Each leaf in the revealed tree is an exposed node.



Single-Child Fraction

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

$$\Pr[\text{at most } 2 \cdot n \cdot \delta \text{ nodes will be exposed}] = 1 - o(1)$$

Each leaf in the revealed tree is an exposed node.

$$\Pr[\text{the revealed tree will have at most } 2 \cdot n \cdot \delta \text{ leaves}] = 1 - o(1)$$

Single-Child Fraction

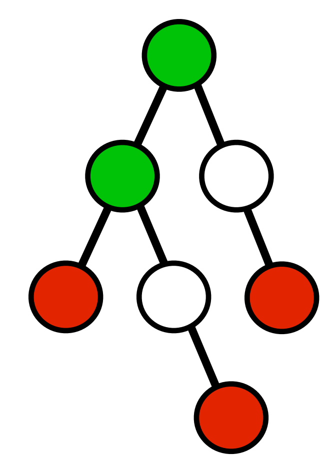
$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

$$\Pr[\text{at most } 2 \cdot n \cdot \delta \text{ nodes will be exposed}] = 1 - o(1)$$

Each leaf in the revealed tree is an exposed node.

$$\Pr[\text{the revealed tree will have at most } 2 \cdot n \cdot \delta \text{ leaves}] = 1 - o(1)$$

In an arbitrary tree, the number of internal nodes with more than one child is upper-bounded by the number of leaves.



Single-Child Fraction

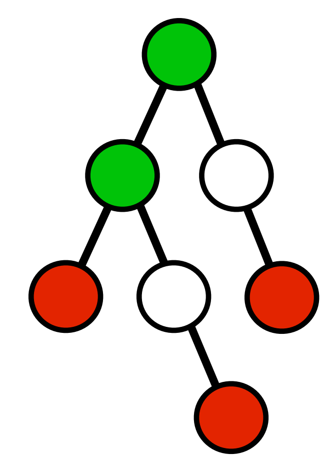
$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

$$\Pr[\text{at most } 2 \cdot n \cdot \delta \text{ nodes will be exposed}] = 1 - o(1)$$

Each leaf in the revealed tree is an exposed node.

$$\Pr[\text{the revealed tree will have at most } 2 \cdot n \cdot \delta \text{ leaves}] = 1 - o(1)$$

In an arbitrary tree, the number of internal nodes with more than one child is upper-bounded by the number of leaves.



$$\Pr[\text{the revealed tree has} \leq 4n\delta \text{ non-single-child nodes}] = 1 - o(1)$$

Single-Child Fraction

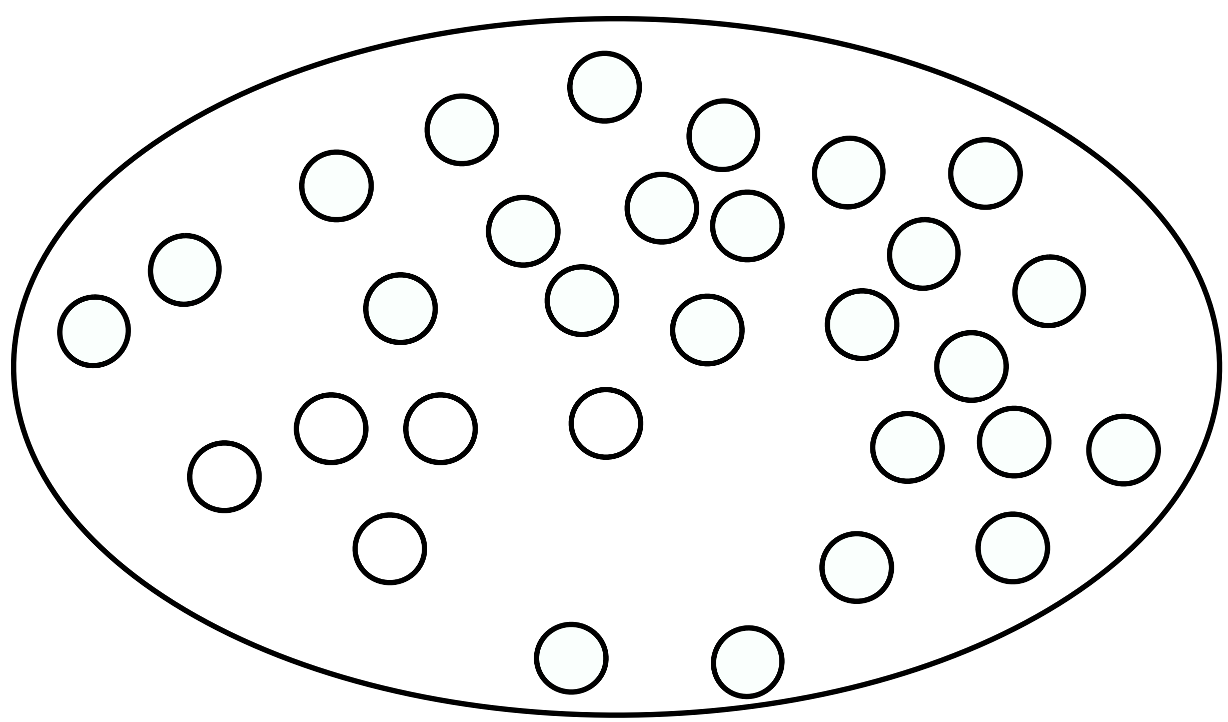
$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

$$\Pr[\text{the revealed tree has} \leq 4n\delta \text{ non-single-child nodes}] = 1 - o(1)$$

Single-Child Fraction

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

$$\Pr[\text{the revealed tree has} \leq 4n\delta \text{ non-single-child nodes}] = 1 - o(1)$$

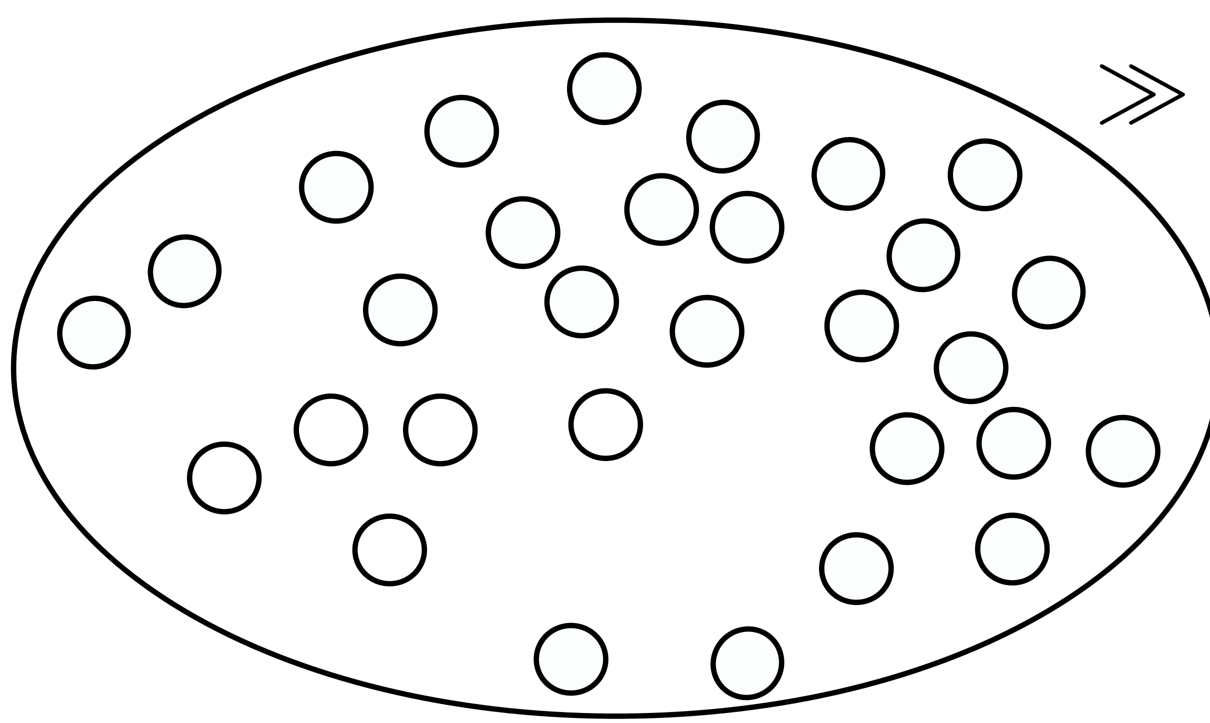


Single-Child Fraction

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

$$\Pr[\text{the revealed tree has} \leq 4n\delta \text{ non-single-child nodes}] = 1 - o(1)$$

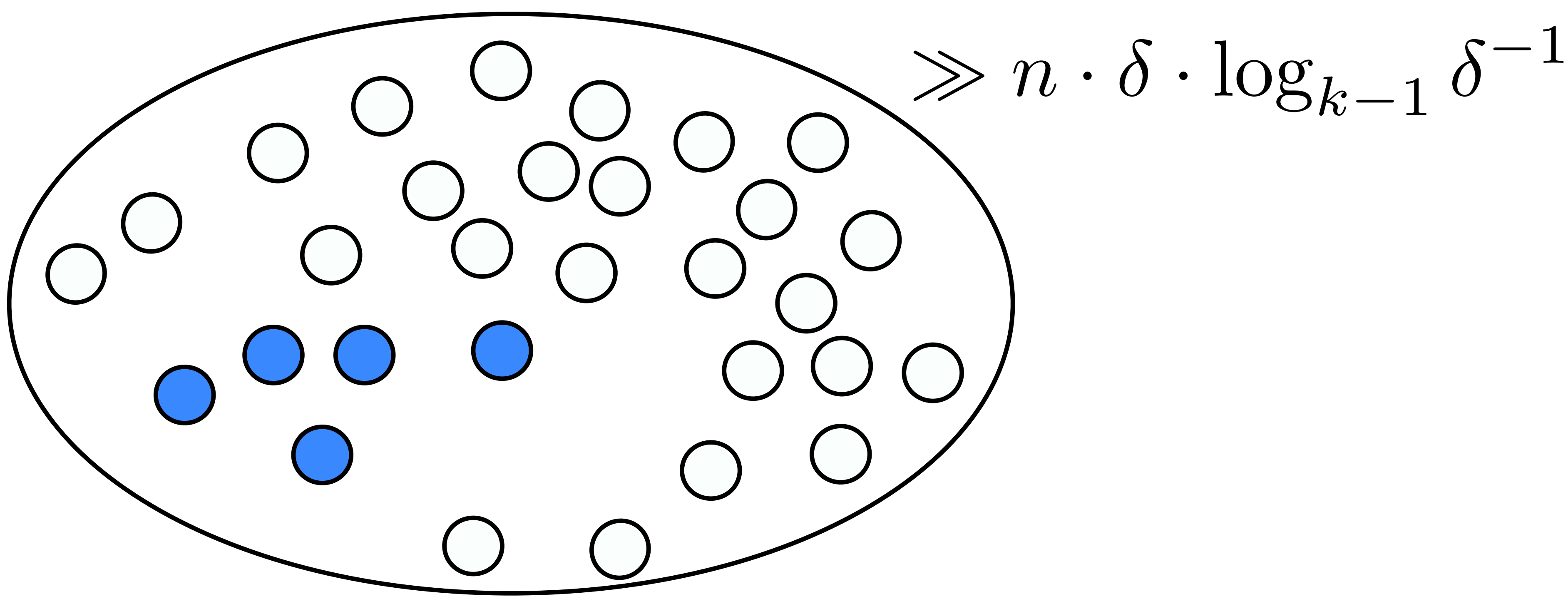
$\gg n \cdot \delta \cdot \log_{k-1} \delta^{-1}$



Single-Child Fraction

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

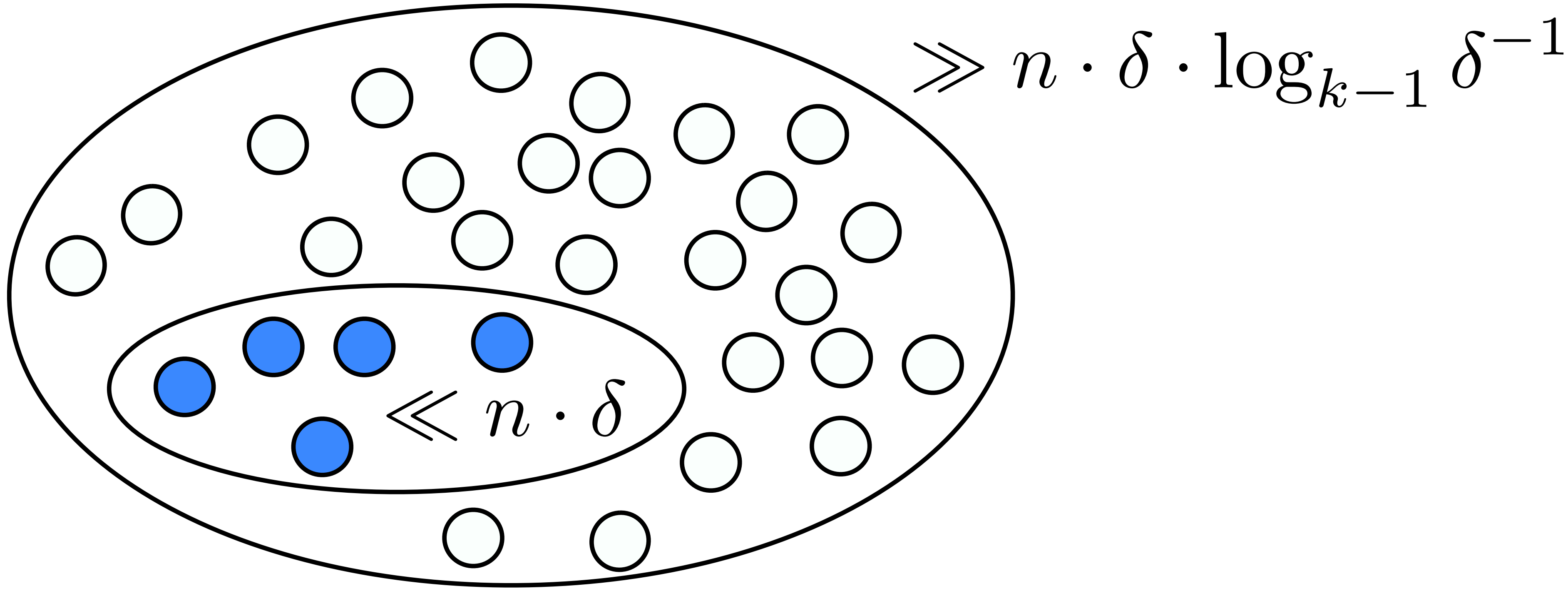
$$\Pr[\text{the revealed tree has} \leq 4n\delta \text{ non-single-child nodes}] = 1 - o(1)$$



Single-Child Fraction

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

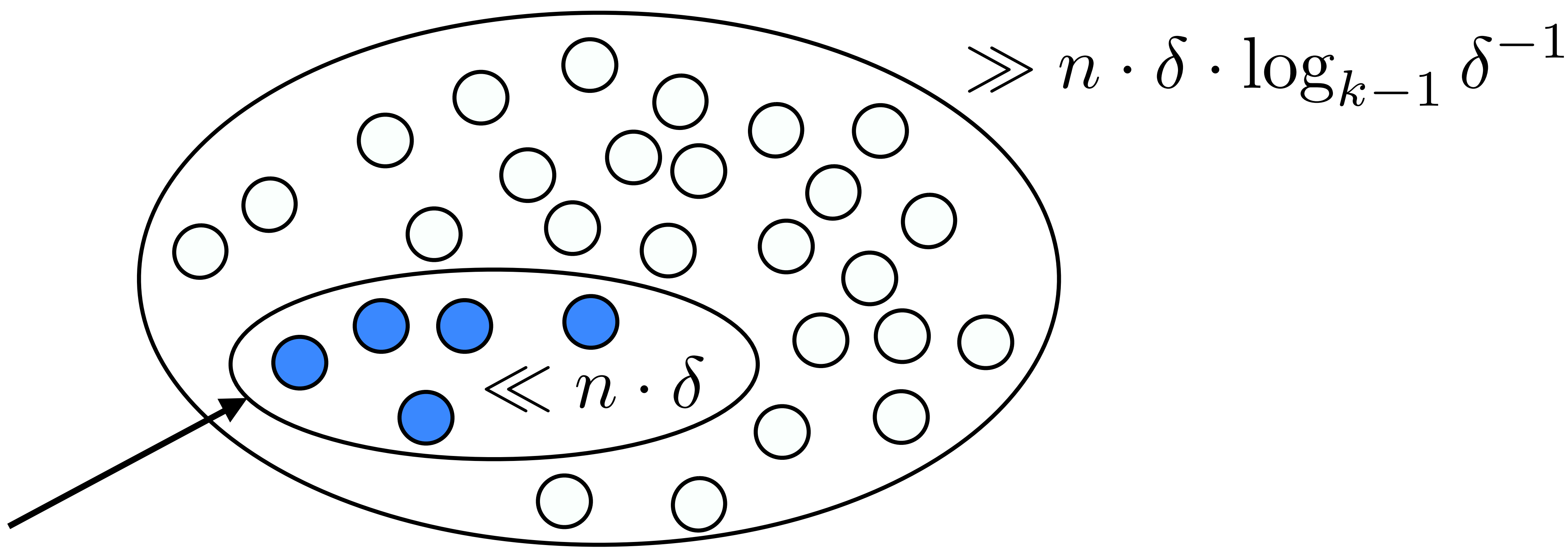
$$\Pr[\text{the revealed tree has} \leq 4n\delta \text{ non-single-child nodes}] = 1 - o(1)$$



Single-Child Fraction

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

$$\Pr[\text{the revealed tree has} \leq 4n\delta \text{ non-single-child nodes}] = 1 - o(1)$$



A $\dfrac{1}{\log_{k-1} \delta^{-1}}$ fraction of the set.

Single-Child Fraction

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

$$\Pr[\text{the revealed tree has} \leq 4n\delta \text{ non-single-child nodes}] = 1 - o(1)$$

$$\Pr\left[\text{the fraction of single-child nodes in the revealed tree is } \geq 1 - O\left(\frac{1}{\log_{k-1} \delta^{-1}}\right)\right] = 1 - o(1)$$

Single-Child Fraction

$$\Pr[\Omega(n \cdot \delta \cdot \log_{k-1} \delta^{-1}) \text{ nodes will be revealed}] = 1 - o(1)$$

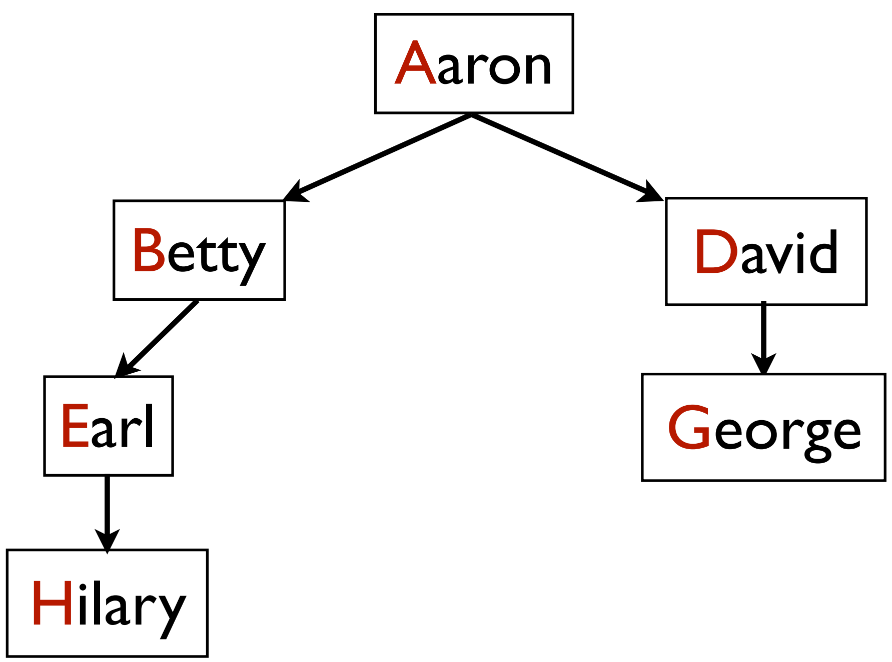
$$\Pr[\text{the revealed tree has} \leq 4n\delta \text{ non-single-child nodes}] = 1 - o(1)$$

$$\Pr\left[\text{the fraction of single-child nodes in the revealed tree is } \geq 1 - O\left(\frac{1}{\log_{k-1} \delta^{-1}}\right)\right] = 1 - o(1)$$

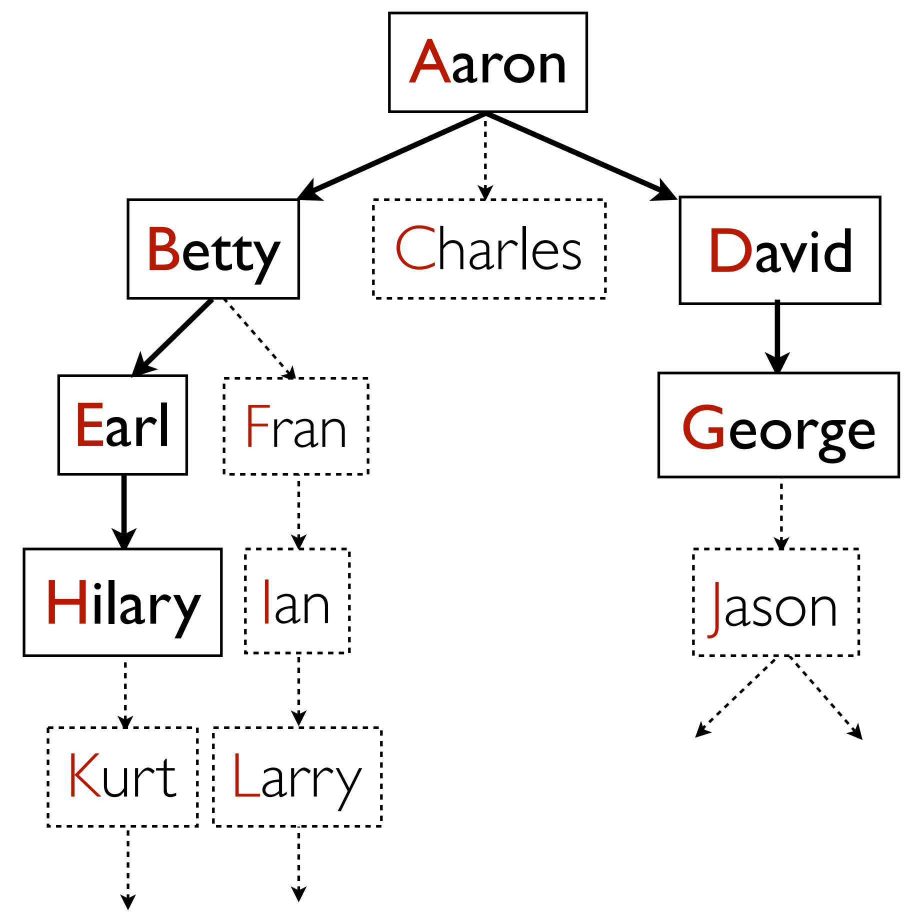
The high single-child fraction can be explained by assuming just a degree bound on the unknown tree

Number of Signers

Revealed Tree

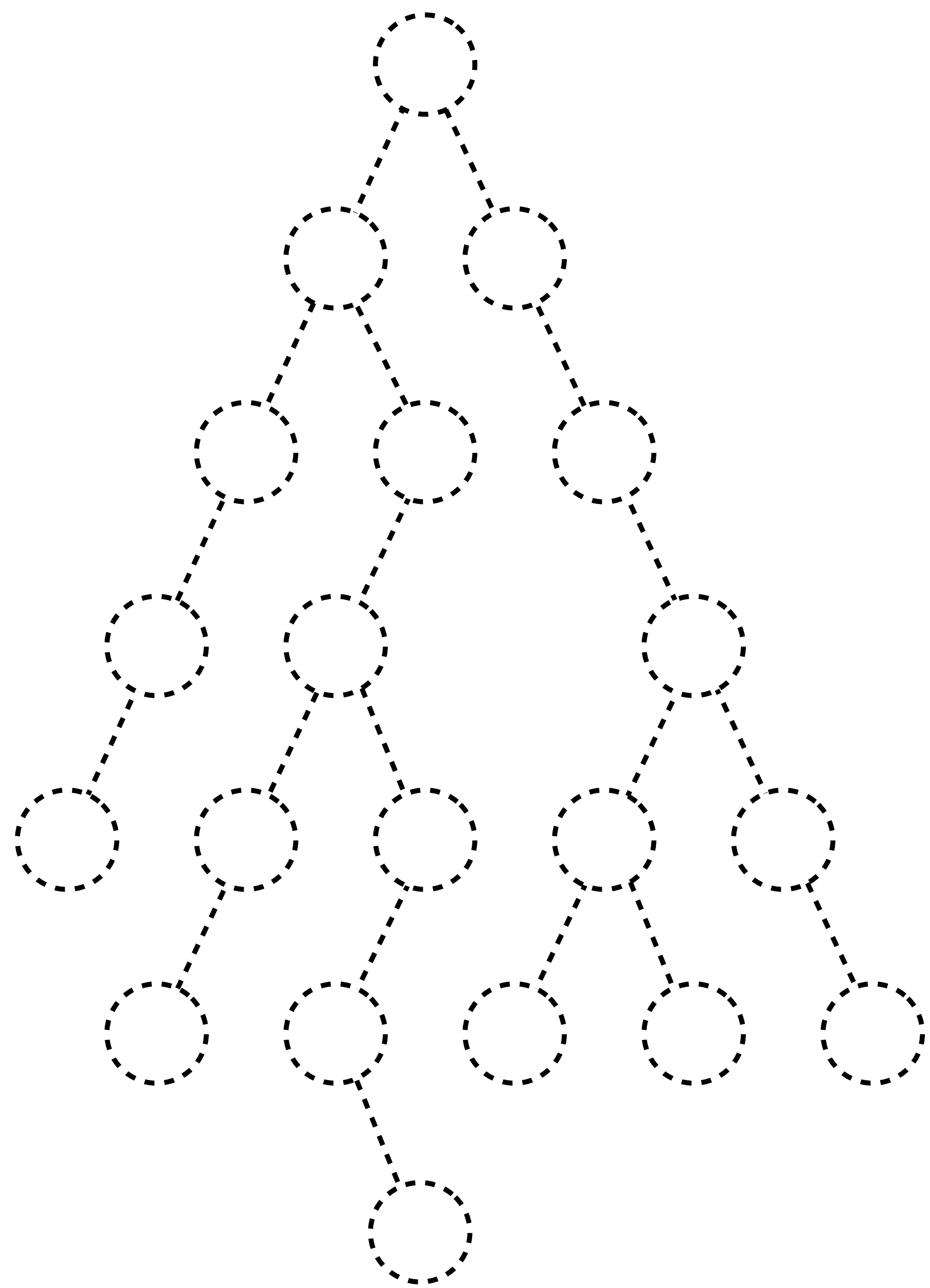


Unknown Tree

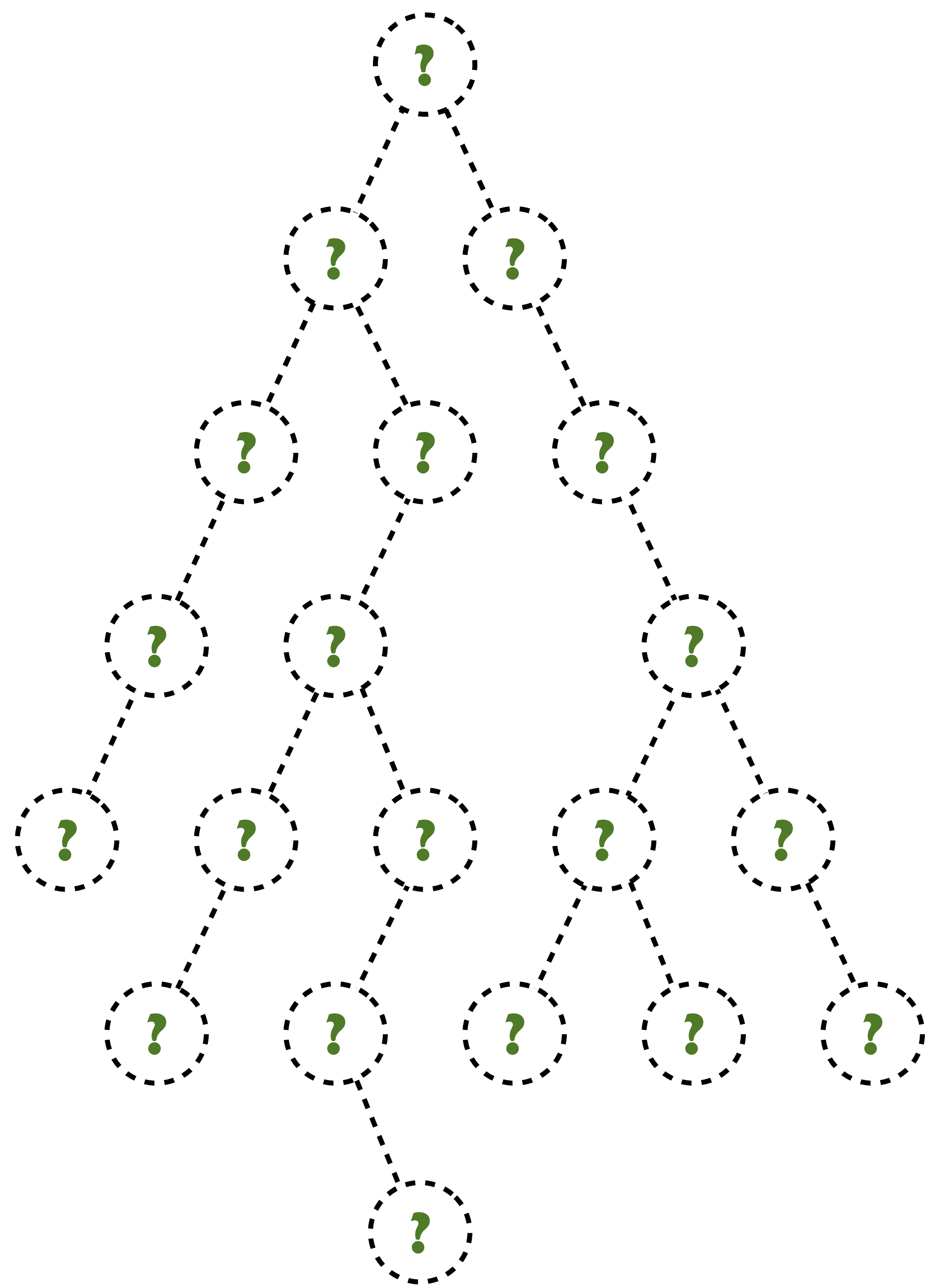


How to guess the size of the unknown tree?

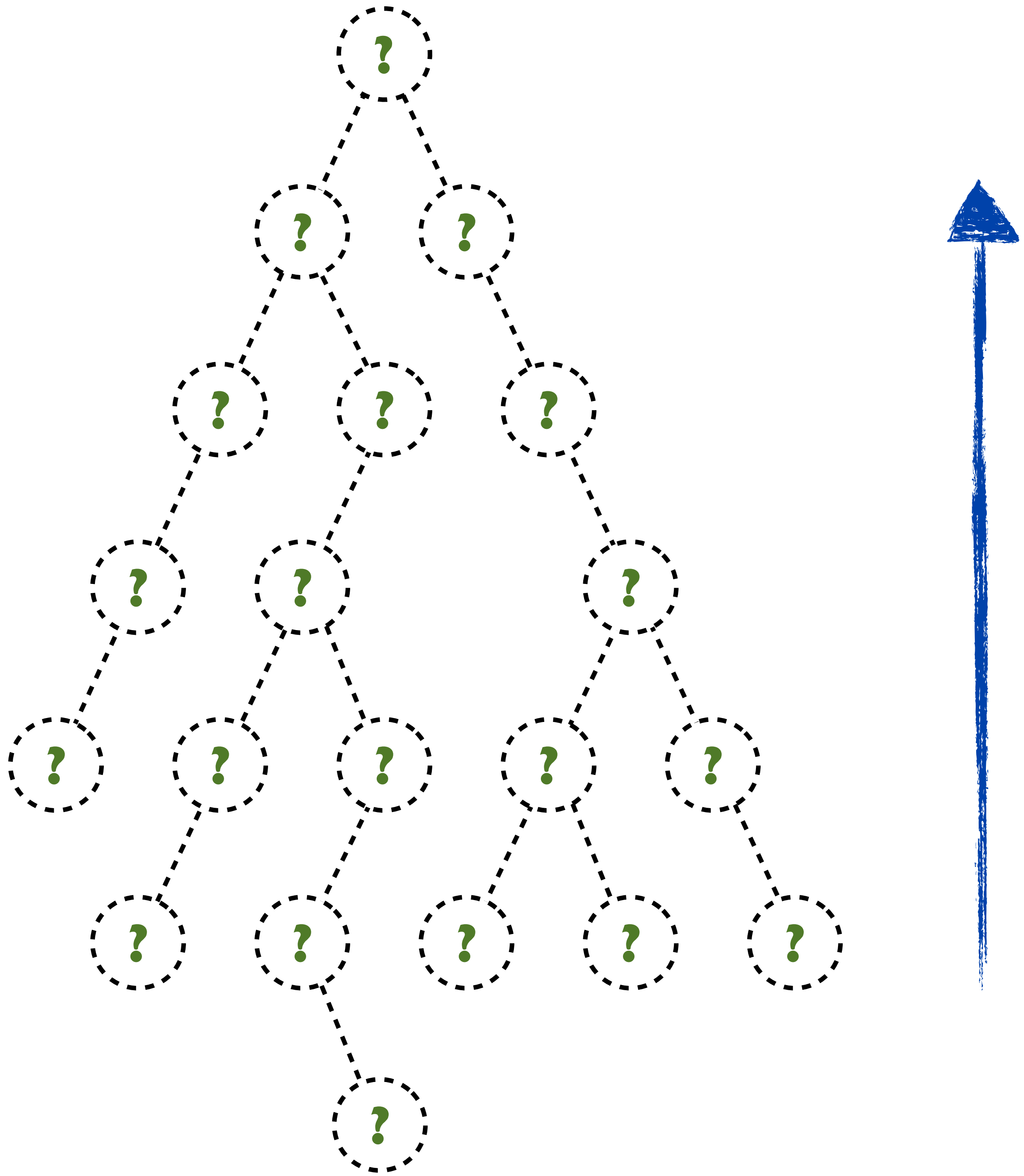
Unknown Tree Exposure



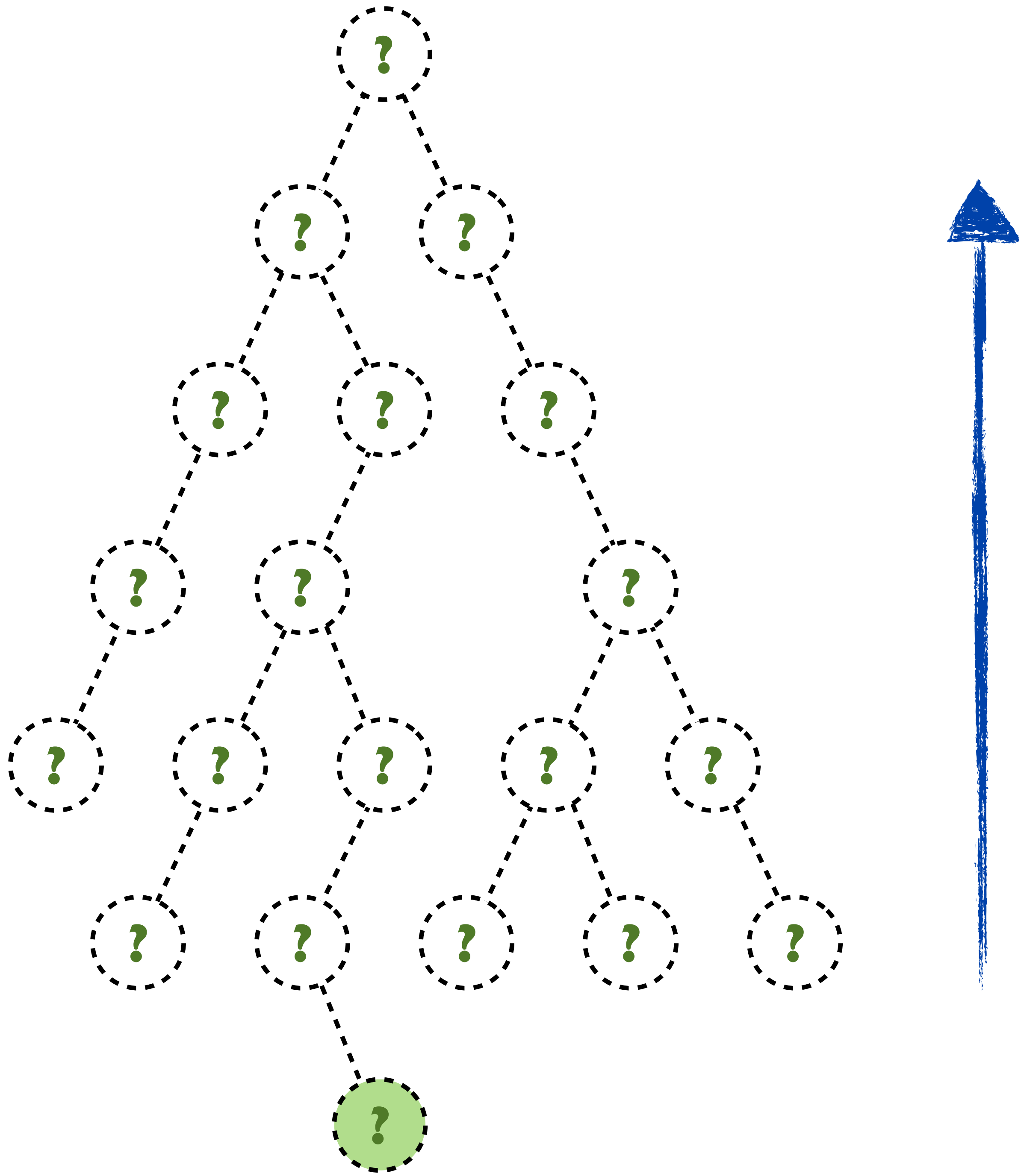
Unknown Tree Exposure



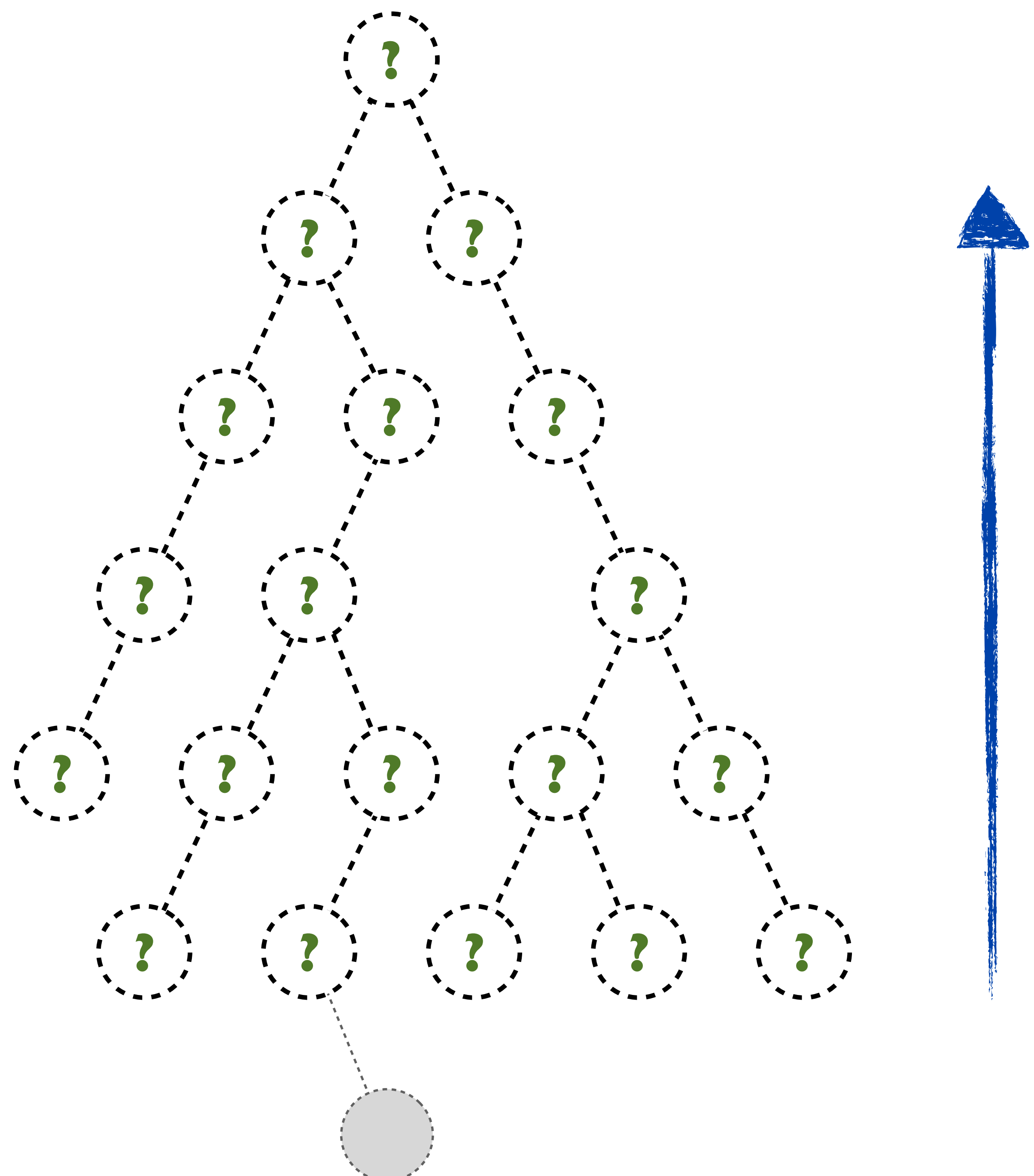
Unknown Tree Exposure



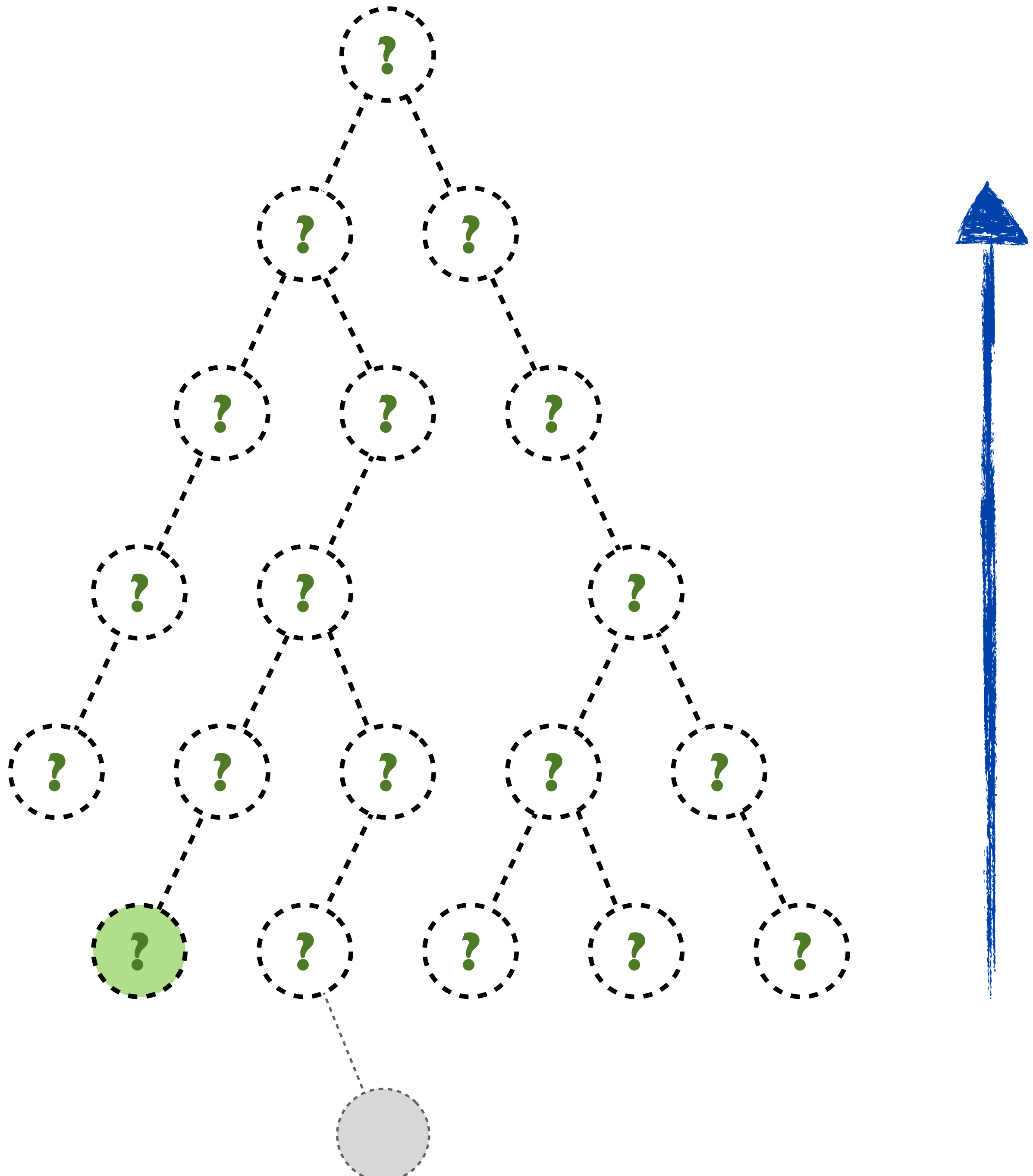
Unknown Tree Exposure



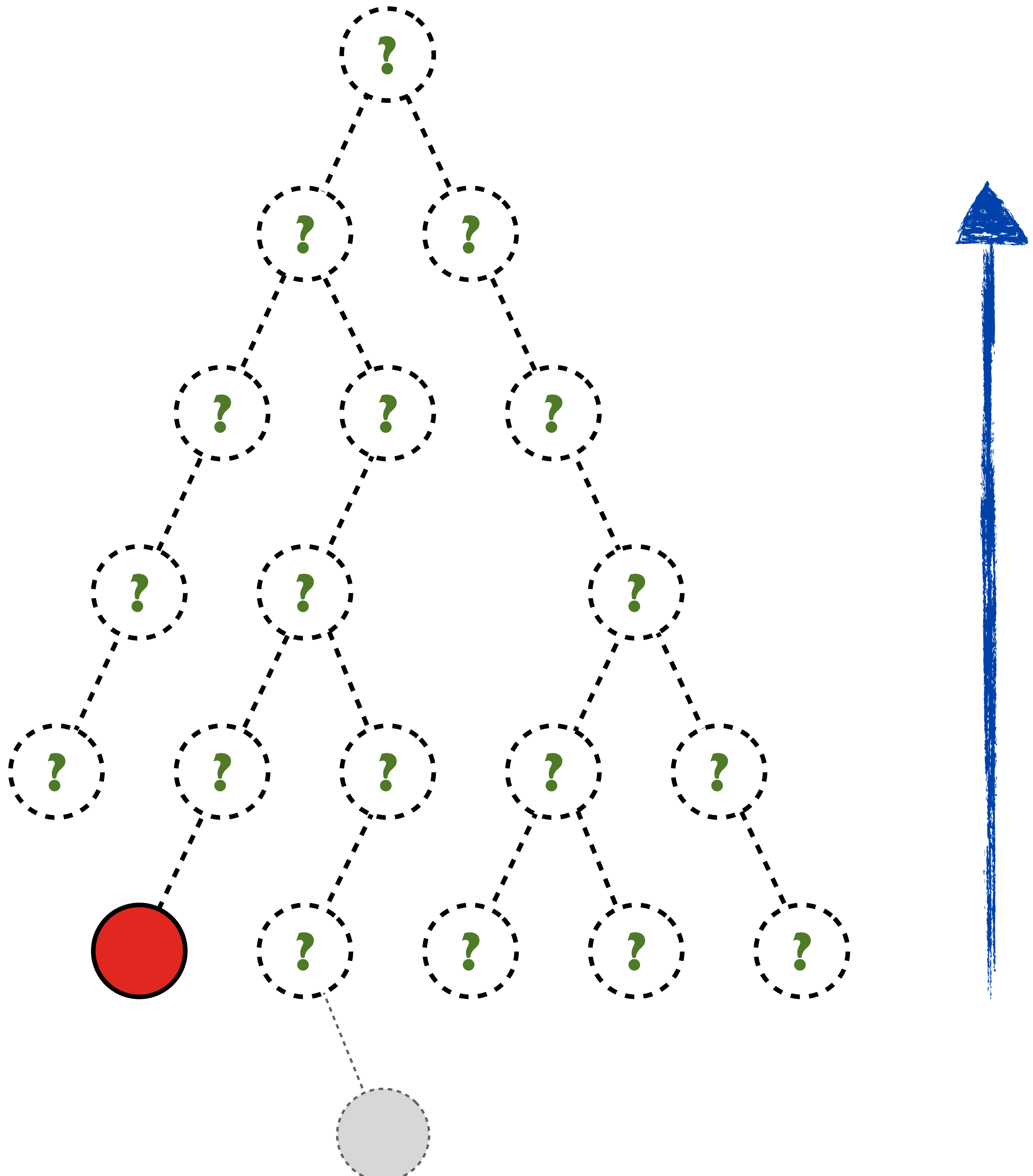
Unknown Tree Exposure



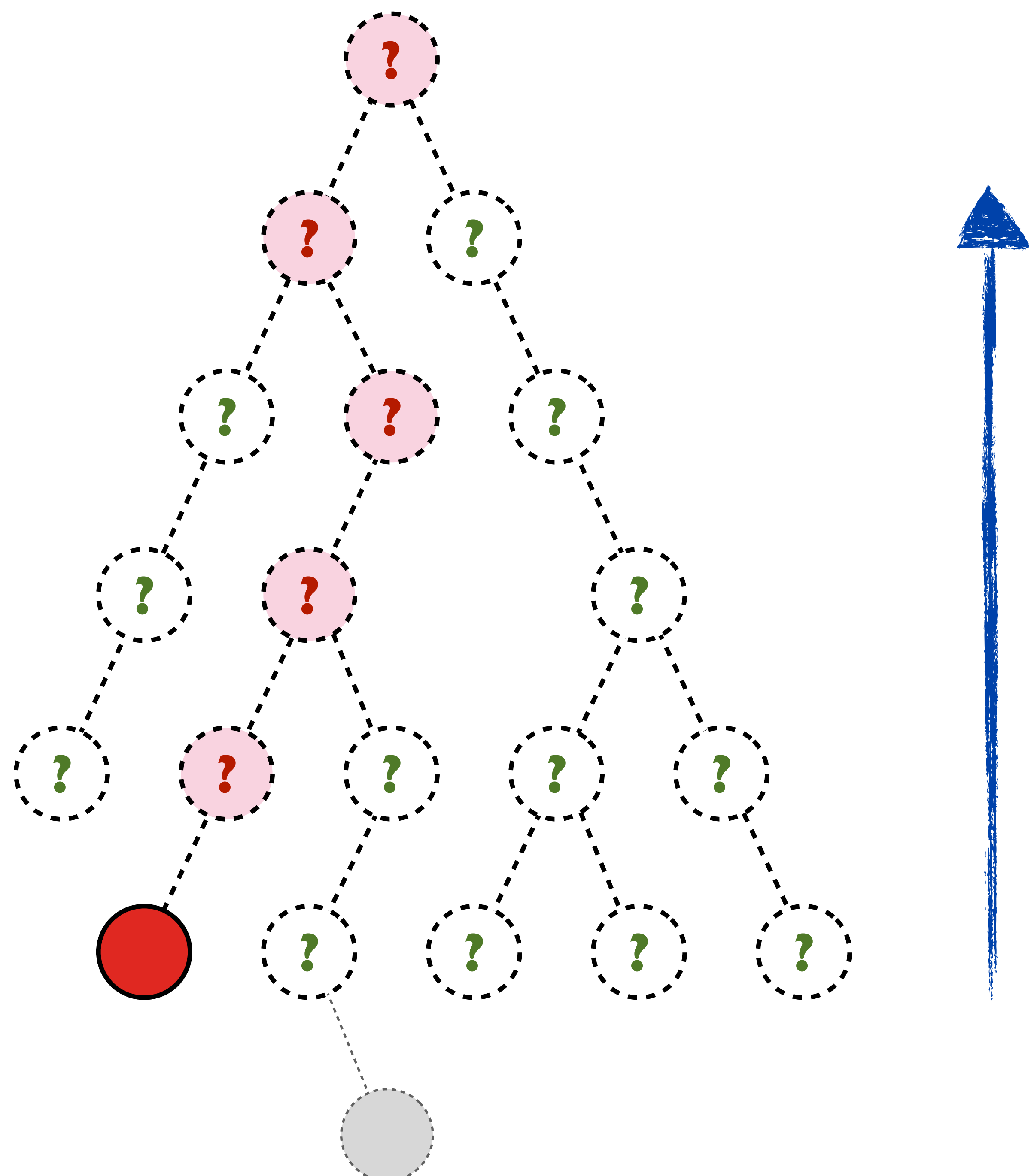
Unknown Tree Exposure



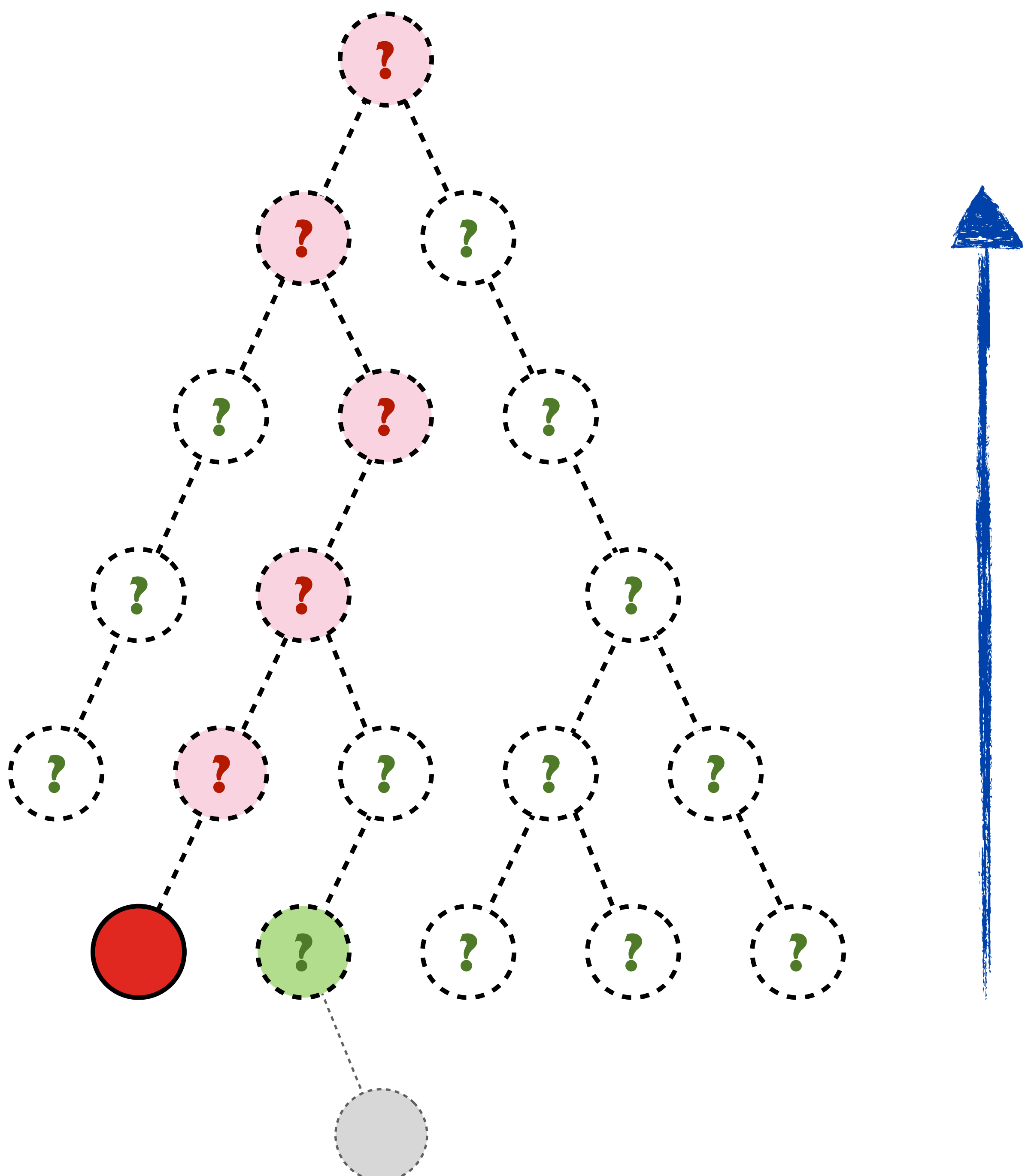
Unknown Tree Exposure



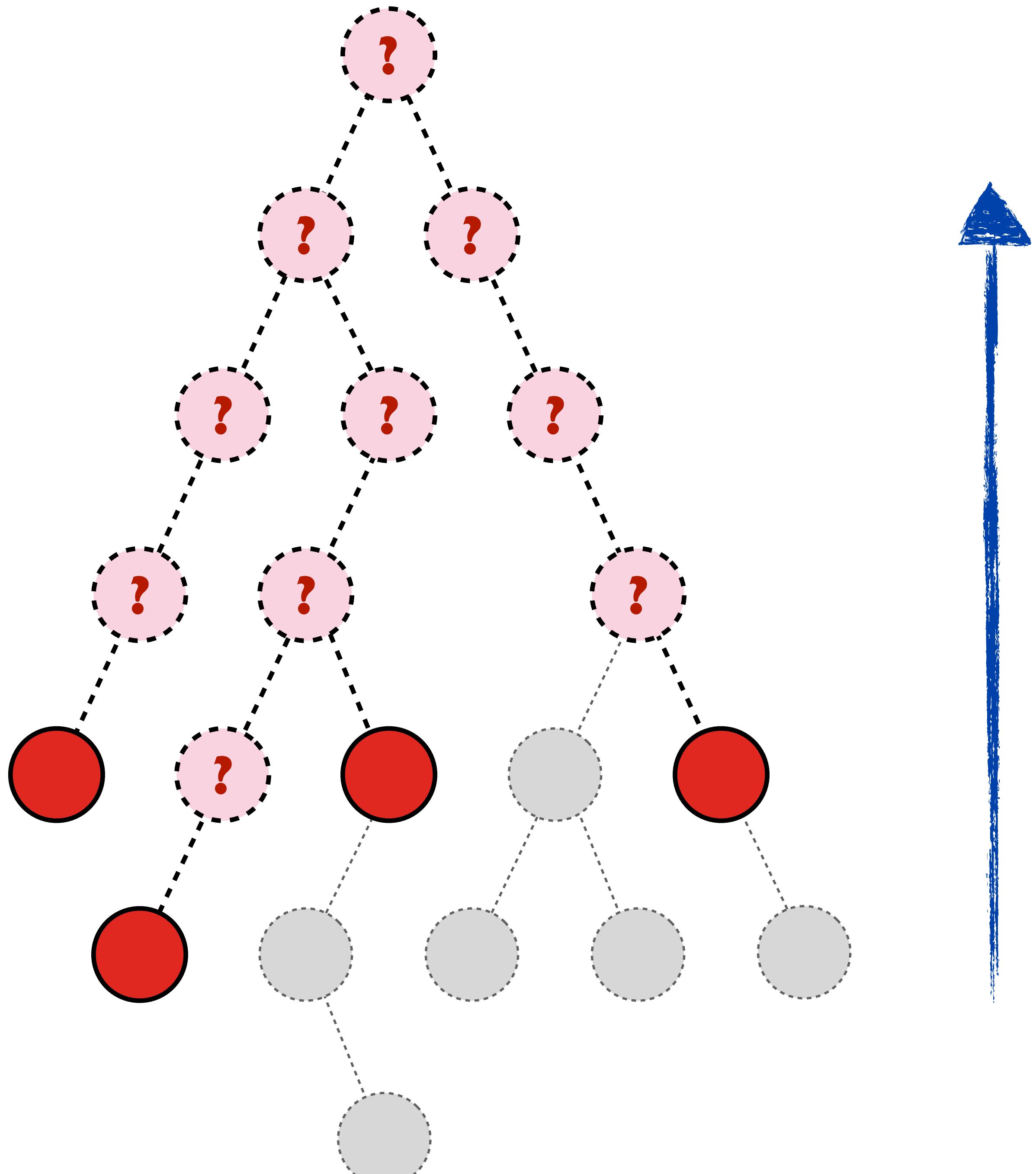
Unknown Tree Exposure



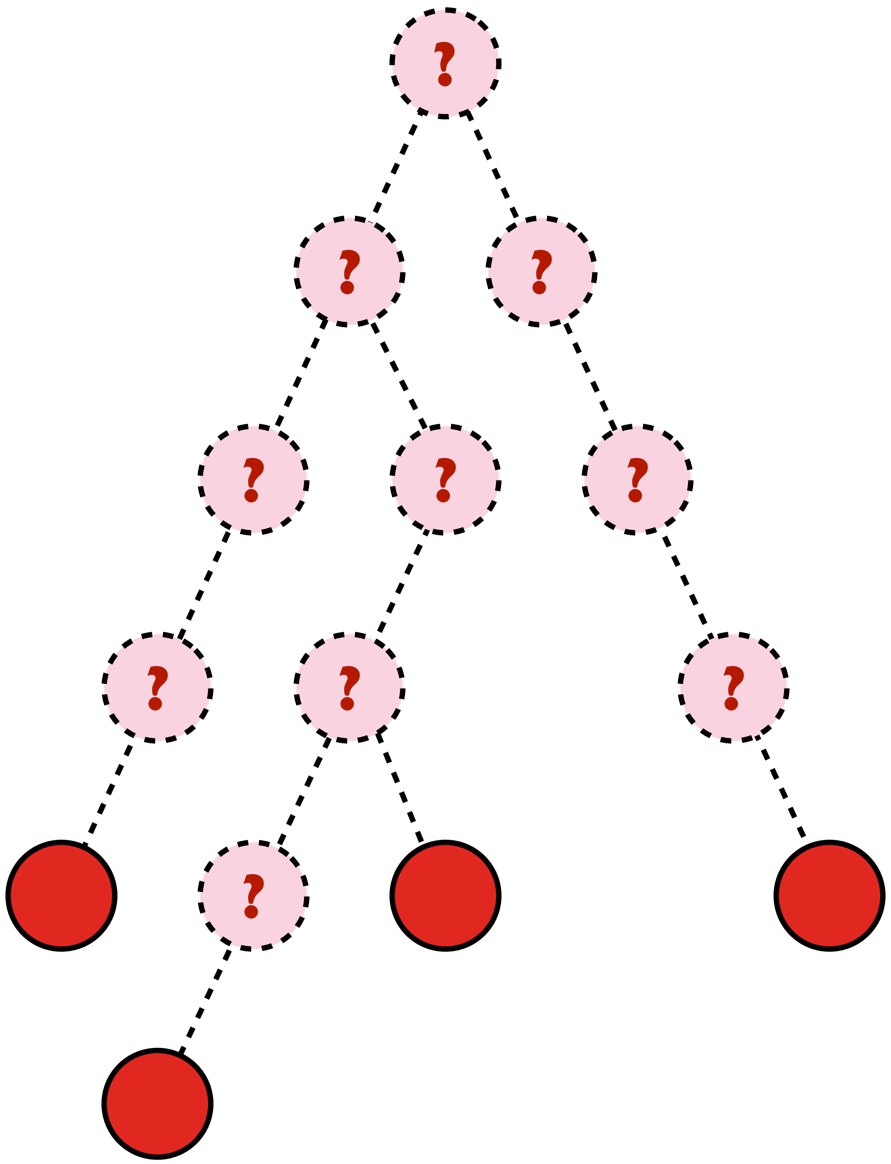
Unknown Tree Exposure



Unknown Tree Exposure

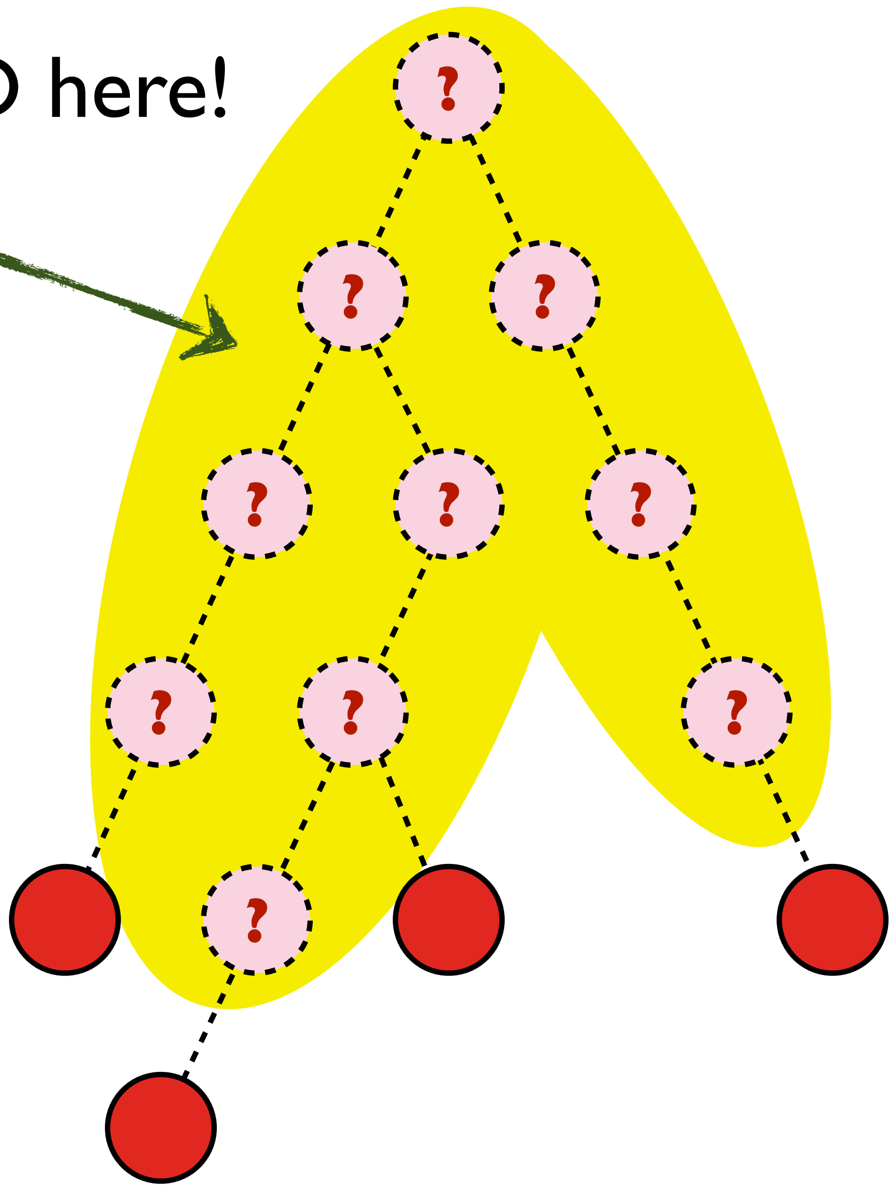


Revealed Tree



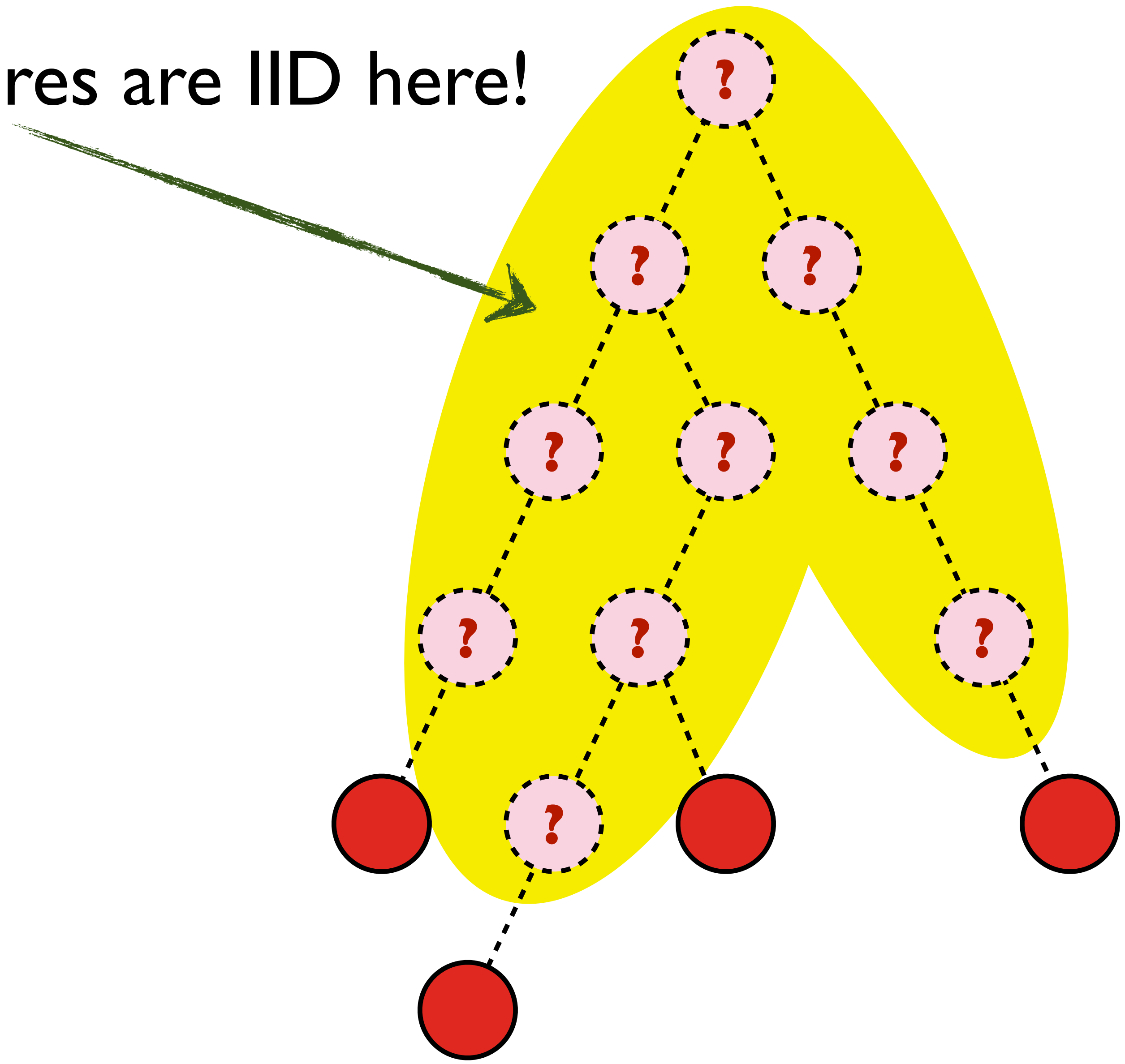
Revealed Tree

Nodes exposures are IID here!



Size Estimation

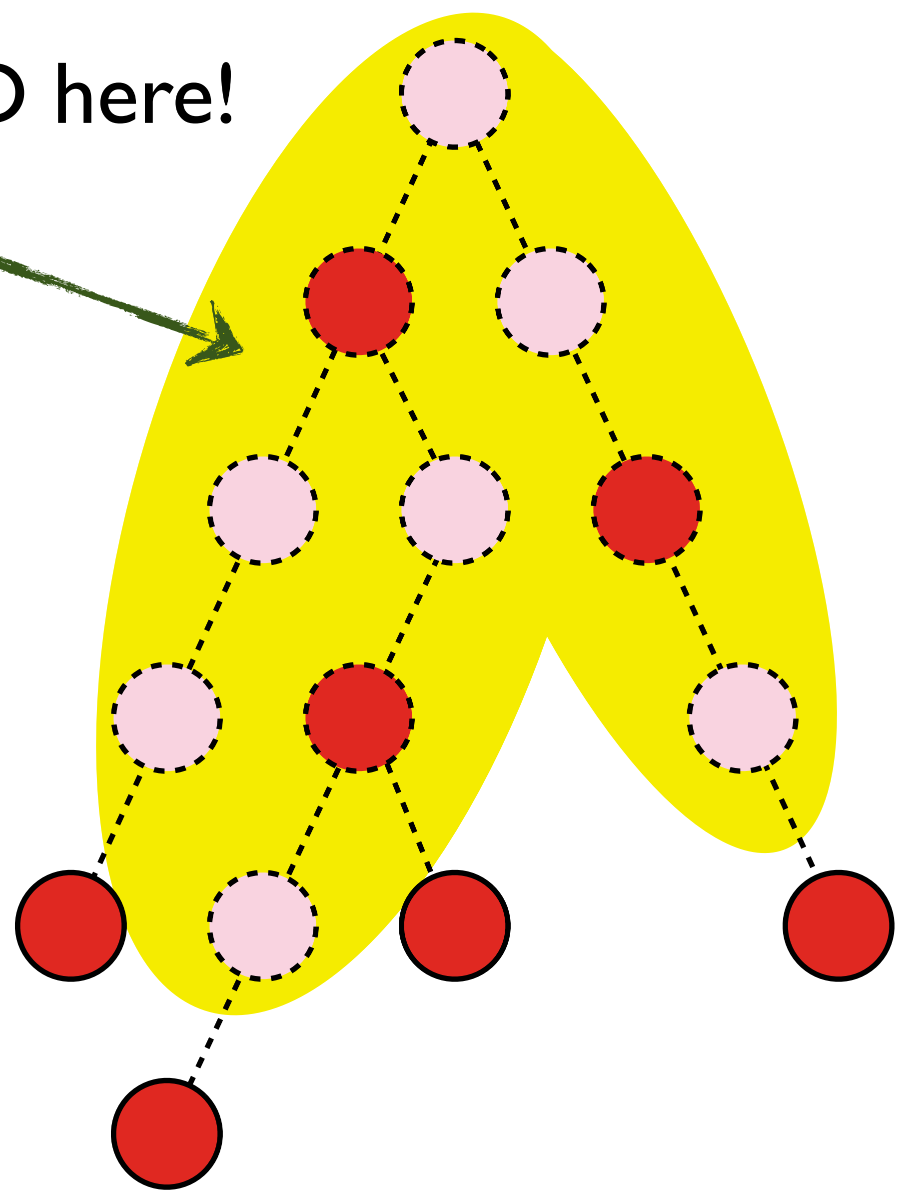
Nodes exposures are IID here!



1. Estimate δ

Size Estimation

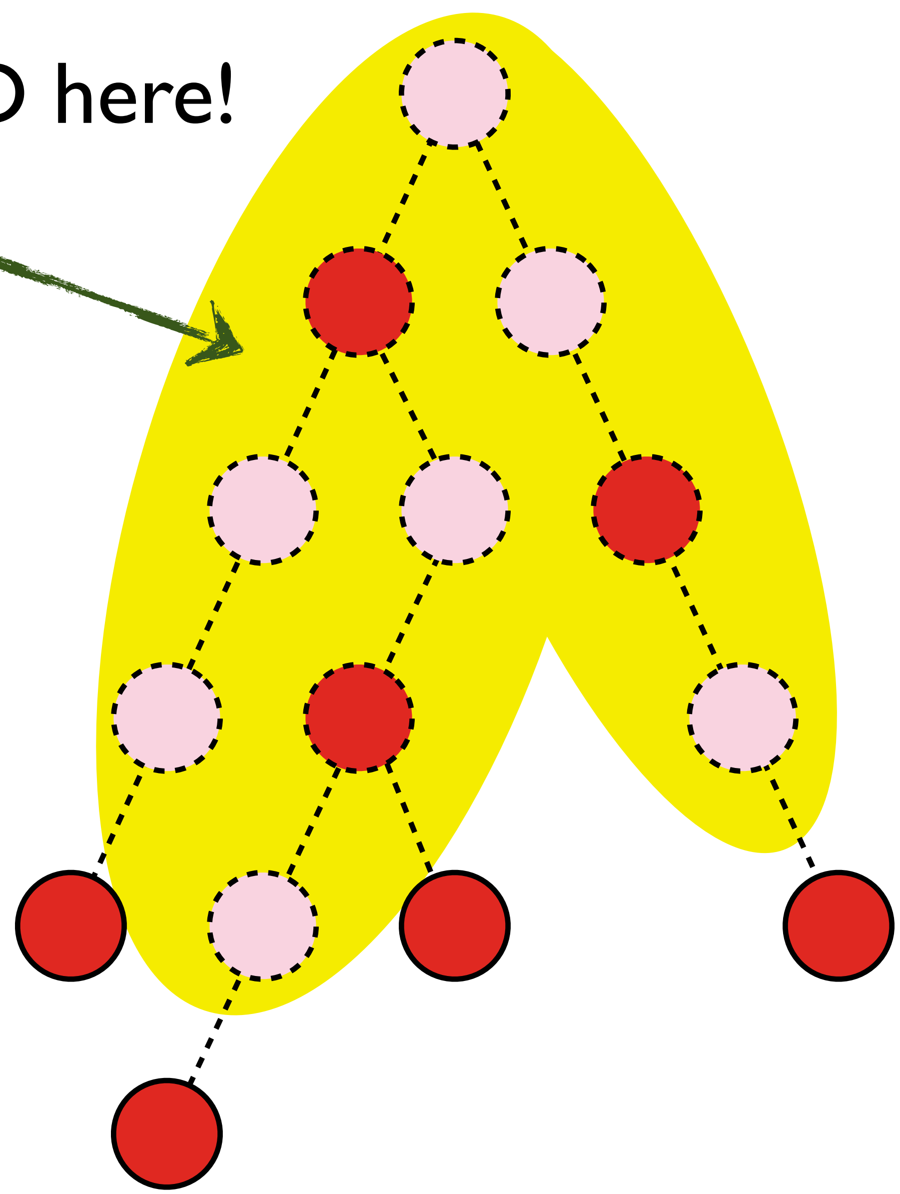
Nodes exposures are IID here!



1. Estimate δ

Size Estimation

Nodes exposures are IID here!

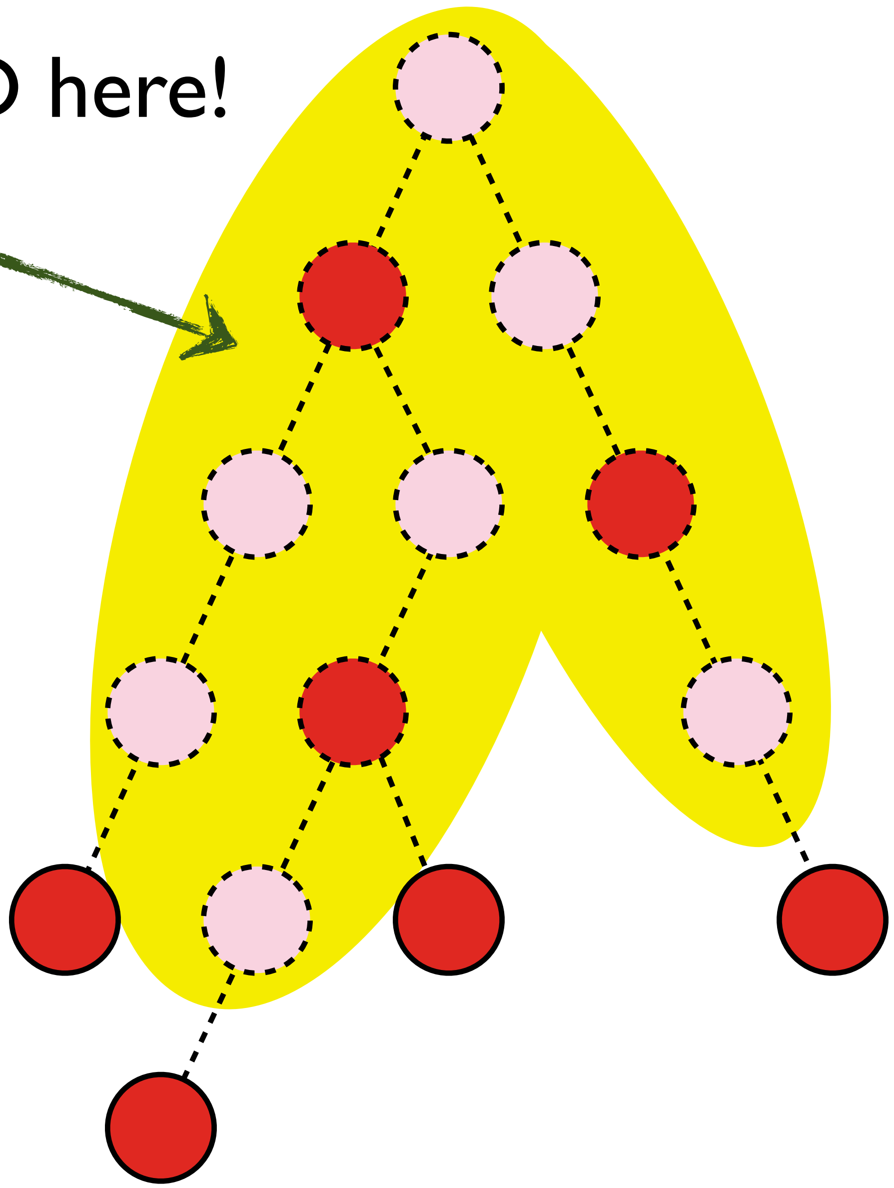


$$\delta \simeq \frac{3}{10}$$

1. Estimate δ

Size Estimation

Nodes exposures are IID here!

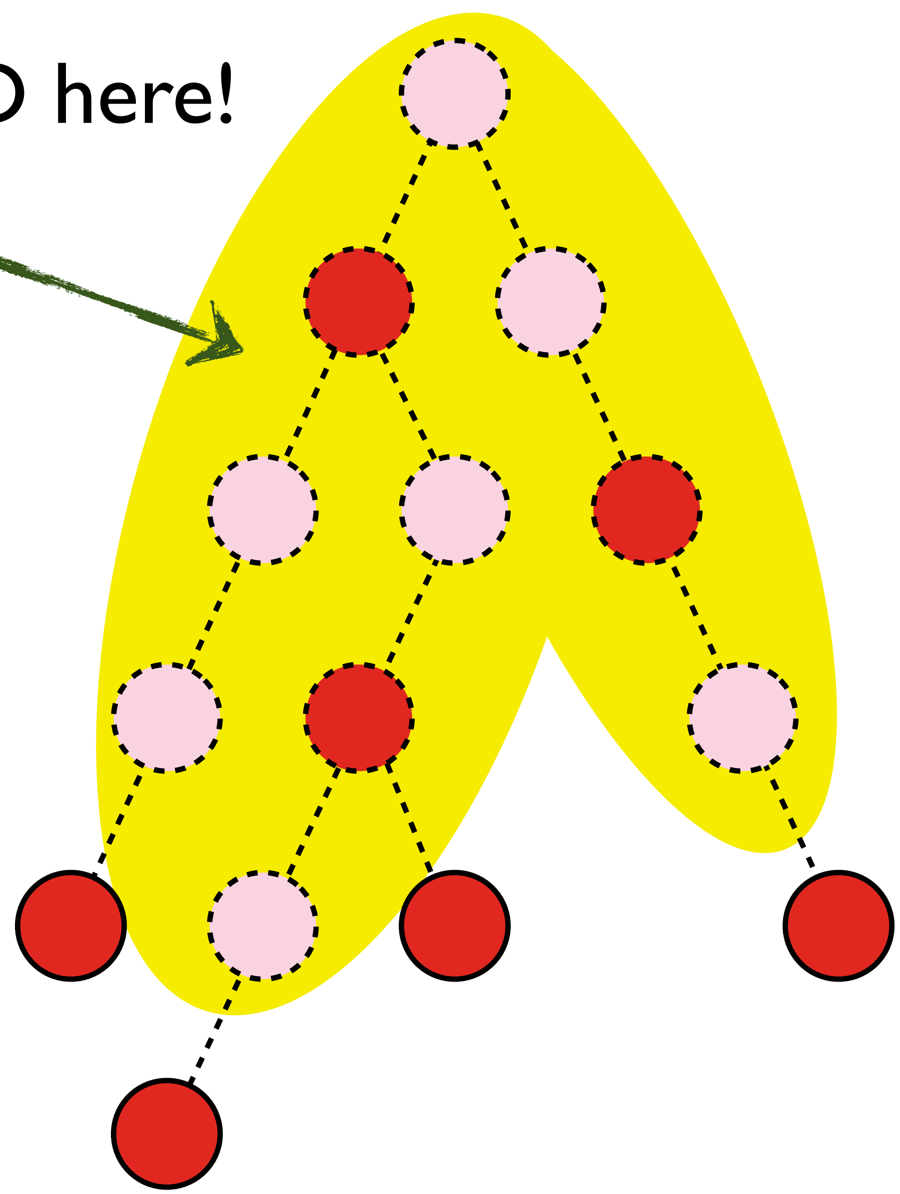


$\delta \simeq \frac{3}{10}$

2. Estimate $n \cdot \delta$ using the number of exposed nodes in the revealed tree

Size Estimation

Nodes exposures are IID here!



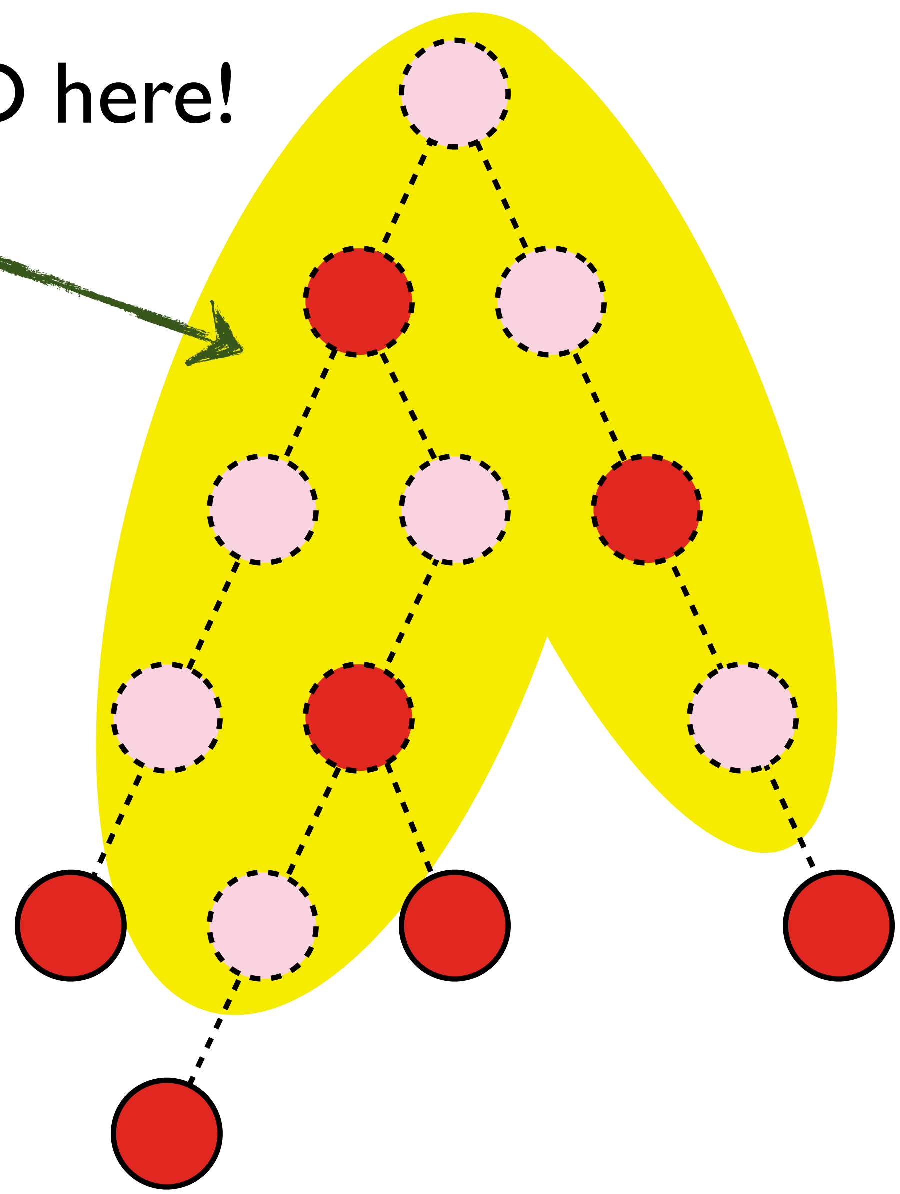
$$\delta \simeq \frac{3}{10}$$

$$n \cdot \delta \simeq 7$$

2. Estimate $n \cdot \delta$ using the number of exposed nodes in the revealed tree

Size Estimation

Nodes exposures are IID here!



$$\delta \simeq \frac{3}{10}$$

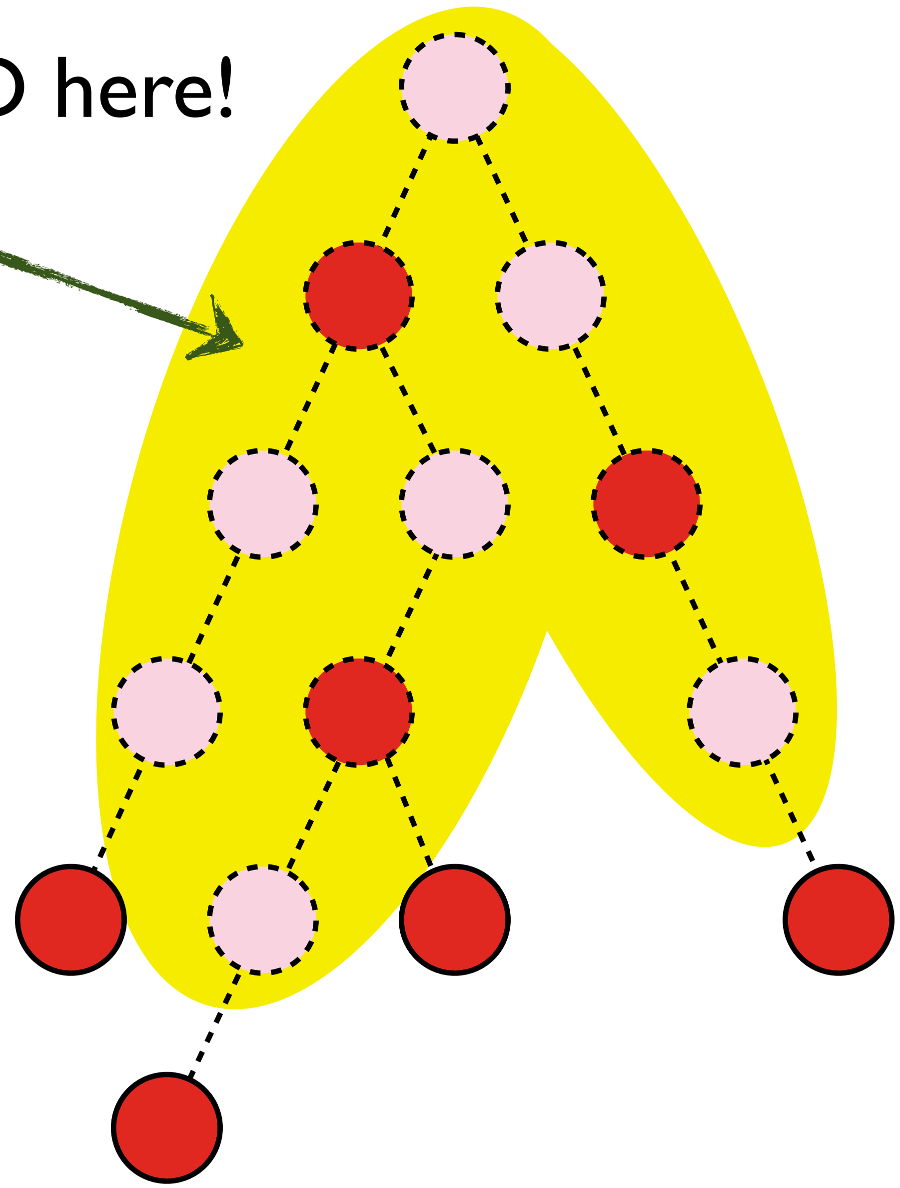
$$n \cdot \delta \simeq 7$$

$$n \simeq 23.\bar{3}$$

3. Take the ratio

Size Estimation

Nodes exposures are IID here!



$$\delta \simeq \frac{3}{10}$$

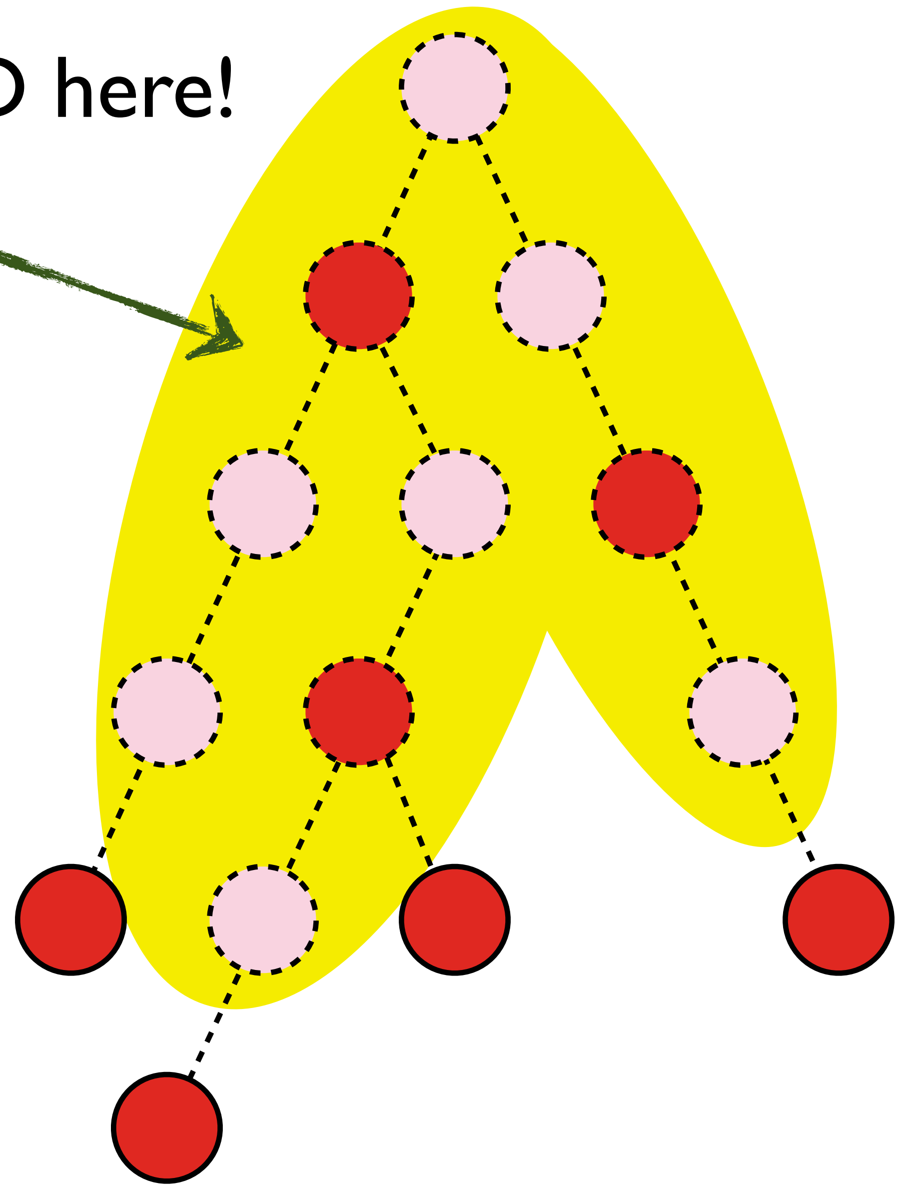
$$n \cdot \delta \simeq 7$$

$$n \simeq 23.\bar{3}$$

What can go wrong?

Size Estimation

Nodes exposures are IID here!



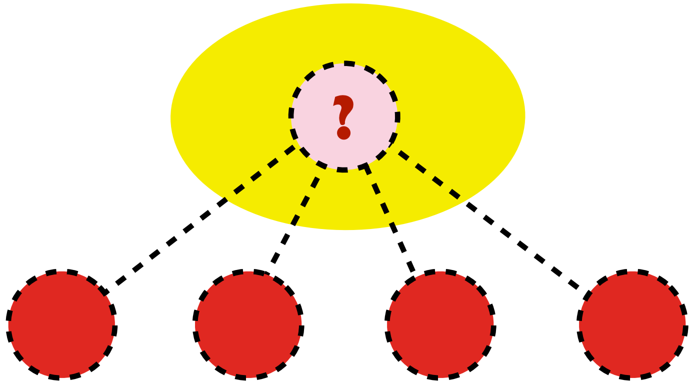
$$\delta \simeq \frac{3}{10}$$

$$n \cdot \delta \simeq 7$$

$$n \simeq 23.\bar{3}$$

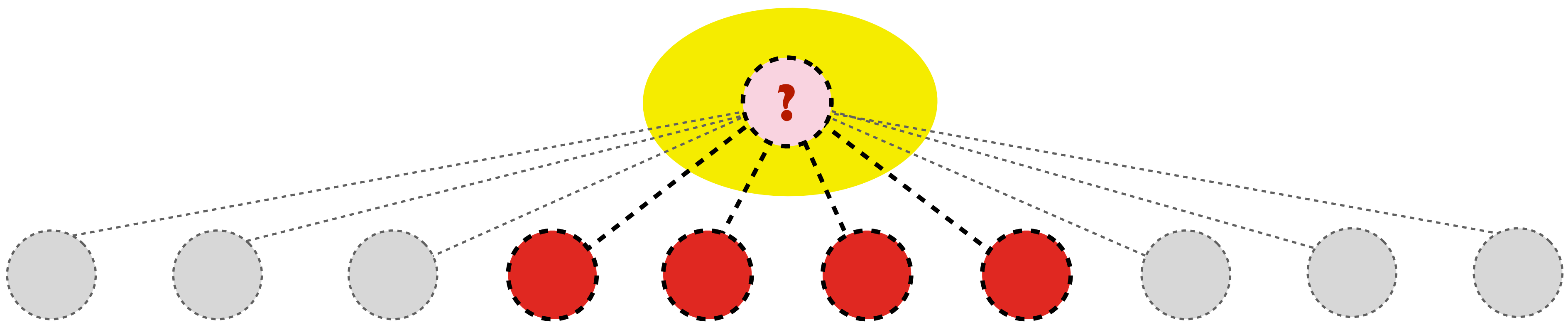
The "yellow area" could contain too few nodes for the estimation of δ to be successful.

Size Estimation



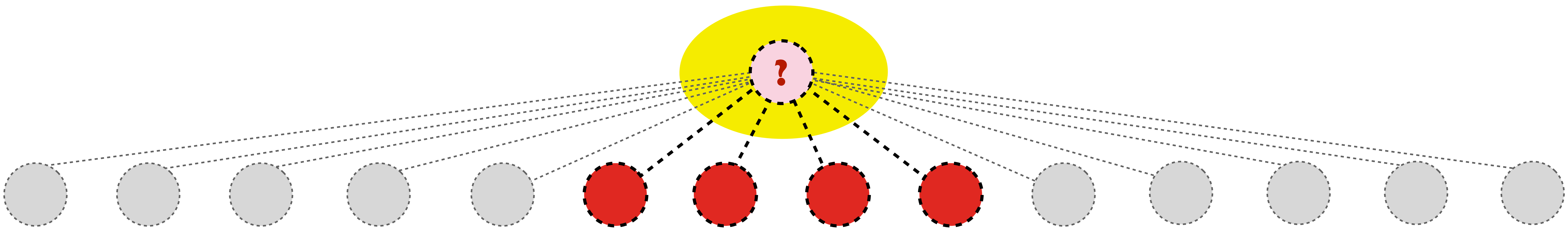
The “yellow area” could contain too few nodes for the estimation of δ to be successful.

Size Estimation



The "yellow area" could contain too few nodes for the estimation of δ to be successful.

Size Estimation



The "yellow area" could contain too few nodes for the estimation of δ to be successful.

Theorem

- The previous algorithm can guess the size with high probability if

$$n > \tilde{\Omega}\left(\max\left(\delta^{-2}, \delta^{-1} \cdot k\right)\right)$$

k is the maximum number of children in the unknown tree,
δ is the exposing probability.

Theorem

- The previous algorithm can guess the size with high probability if

$$n > \tilde{\Omega}\left(\max\left(\delta^{-2}, \delta^{-1} \cdot k\right)\right)$$

k is the maximum number of children in the unknown tree,

δ is the exposing probability.

$k < \tilde{O}\left(\sqrt{n}\right)$ $\delta > \tilde{\Omega}\left(\sqrt{\frac{1}{n}}\right)$ satisfy the requirement

Theorem

- The previous algorithm can guess the size with high probability if

$$n > \tilde{\Omega}\left(\max\left(\delta^{-2}, \delta^{-1} \cdot k\right)\right)$$

  k is the maximum number of children in the unknown tree,
  δ is the exposing probability.

- No algorithm can do it if n is smaller.

IRAQ Tree Size

- We refined our asymptotic theorem for the IRAQ revealed tree (18k nodes)
- Assuming the tree-revealing model,
  we estimate that
  the number of signers of the IRAQ petition
  is within a factor of 2 of 173k
  with probability $\geq 95\%$

Conclusion

Reconstruction Problems

- Answers depend on:
  - the random model that drives the process;
  - the number of traces we have;
  - the random algorithm that reveals pieces of information by “cutting” them out of a model run.

Random Model

- It is important to check whether the random model is close to reality;

- without this step, the guesses might be “close” to the model’s runs, but very far from reality.

This step is possibly the hardest one to do.

Random Model

- It is important to check whether the random model is close to reality;

- without this step, the guesses might be “close” to the model’s runs, but very far from reality.

This step is possibly the hardest one to do.

Checking whether the guesses
are close to reality is sometimes
the only viable approach.

Number of Traces

- If we want to get a significant understanding of some partially-known process,

- we first and foremost need to verify whether the questions we are asking can be significantly answered by the amount of data we have.

This step can be carried out
(theorems and/or simulations)

Trace Generation

- The number of “traces” needed depends strongly on which pieces of information of a model’s run are revealed (as well as on the random model).

- In some cases, it might be possible to get more informative traces by a deeper mining of the data.

Complexity of Guessing

- **If not enough traces are available, we should simplify the questions** we ask about the unknown process.

- Otherwise, our guesses might be completely off and insignificant.

Complexity of Guessing

- How many traces do we need for other guessing tasks?
- Very, very, rich area of problems.

Thanks!