Reconstructing Patterns of Information Diffusion from Incomplete Observations

Flavio Chierichetti  
Jon Kleinberg  
David Liben-Nowell  

Cornell University  
Cornell University  
Carleton College
Internet Activism

• Very important phenomenon.
• Incomplete Traces:
  how to study partially-visible viral phenomena?
• Chain Letter Petitions:
  how to estimate the reach?
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
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
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
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

Aaron

Betty  Charles  David
Chain Letters

Aaron

Betty

Charles

David

Earl

Fran
Chain Letters

Aaron

Betty

Charles

David

Earl

Fran

George

Hilary
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

The signers,
Betty
Charles
David
Fran
George
Hilary

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

Aaron

Betty

Charles

David

Earl

Fran

George

Hilary
Chain Letters

Aaron

Betty

Charles

David

Earl
Fran

Hilary

George
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,
[Names listed]

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,
[Names listed]
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

Aaron

Betty

Charles

David

Earl Fran

Hilary

George
Chain Letters

- Aaron
  - Betty
    - Earl
    - Fran
    - Hilary
  - Charles
  - David
    - George
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
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
...
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

- Aaron
  - Betty
    - Earl
      - Hilary
  - Charles
  - David
    - George
Tree-Revealing Process

Liben-Nowell, Kleinberg, PNAS’08

Each node is exposed independently with prob. $\delta > 0$
Each node is exposed independently with prob. $\delta > 0$
Tree-Revealing Process

*Liben-Nowell, Kleinberg, PNAS’08*

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

- Aaron
  - Betty
    - Earl
      - Hilary
  - David
    - George
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!

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 $\delta$. 
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
We partition the tree into subforests, in such a way that each subforest has $\sim \delta^{-1}$ 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 $F$
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 $\Omega \left( \log_{k-1} \delta^{-1} \right)$. 
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 \( \Omega \left( \log_{k-1} \delta^{-1} \right) \).
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 $\Omega \left( \log_{k-1} \delta^{-1} \right)$. 
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 \( \Omega \left( \log_{k-1} \delta^{-1} \right) \).

\[ \text{Pr}[\text{some node is exposed in } F's \text{ lower half}] = \Theta(1) \]
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 $\Omega \left( \log_{k-1} \delta^{-1} \right)$. 

Pr[some node is exposed in $F$’s lower half] = $\Theta(1)$
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 \( \Omega \left( \log_{k-1} \delta^{-1} \right) \).

The probability that some node is exposed in \( F \)'s lower half is \( \Theta(1) \).

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

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

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

Nodes exposures are IID here!
Nodes exposures are IID here!

1. Estimate $\delta$
Size Estimation

Nodes exposures are IID here!

1. Estimate $\delta$
Nodes exposures are IID here!

1. Estimate $\delta$

$$\delta \approx \frac{3}{10}$$
Nodes exposures are IID here!

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

\[ \delta \approx \frac{3}{10} \]
Nodes exposures are IID here!

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

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

$$n \cdot \delta \approx 7$$
Size Estimation

Nodes exposures are IID here!

\[ \delta \approx \frac{3}{10} \]

\[ n \cdot \delta \approx 7 \]

\[ n \approx 23.3 \]

3. Take the ratio
Nodes exposures are IID here!

What can go wrong?

\[ \delta \approx \frac{3}{10} \]

\[ n \cdot \delta \approx 7 \]

\[ n \approx 23.3 \]
Nodes exposures are IID here!

The "yellow area" could contain too few nodes for the estimation of $\delta$ to be successful.
Size Estimation

The “yellow area” could contain too few nodes for the estimation of $\delta$ to be successful.
Size Estimation

The “yellow area” could contain too few nodes for the estimation of $\delta$ to be successful.
Theorem

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

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

\( k \) is the maximum number of children in the unknown tree,
\( \delta \) 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,
\(\delta\) is the exposing probability.

\[ k < \tilde{O} \left( \sqrt{n} \right) \quad \delta > \tilde{\Omega} \left( \sqrt{\frac{1}{n}} \right) \]

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,
\( \delta \) 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)
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

• We gave a mathematical explanation of some odd properties observed in real-world revealed trees,

• we used the available revealed trees to guess properties of unknown chain-letter trees.

• We applied our technique to a real-world dataset, giving the first estimate of the number of signers of the IRAQ chain letter.

http://petitions.cs.cornell.edu/