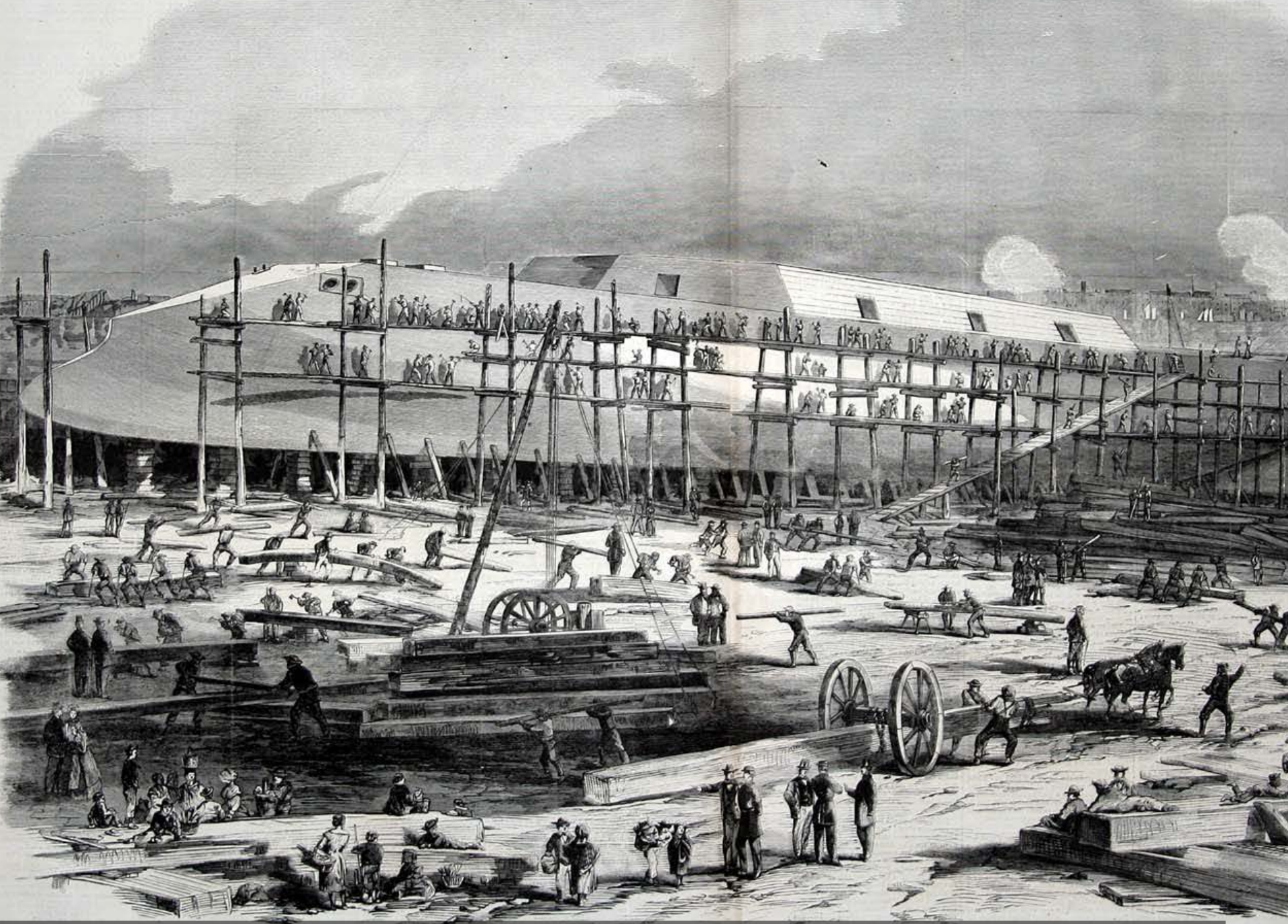
Sampling biases in humanistic networks

{ How systematic biases affect humanities research and why we should care. Also: Hope.

Scott Weingart | Indiana University | @scott_bot | scbweing@indiana.edu
Networks and Network Analysis for the Humanities: Reunion Conference
University of California | Los Angeles | 314 Royce Hall | 9:30-10:00

compiled from visualcomplexity.org



http://www.sonofthesouth.net/leefoundation/civil-war/1863/november/ship-building-1500.jpg

carta marina: http://upload.wikimedia.org/wikipedia/commons/e/e5/Carta_Marina_AB_stitched.jpg

MISTAKES

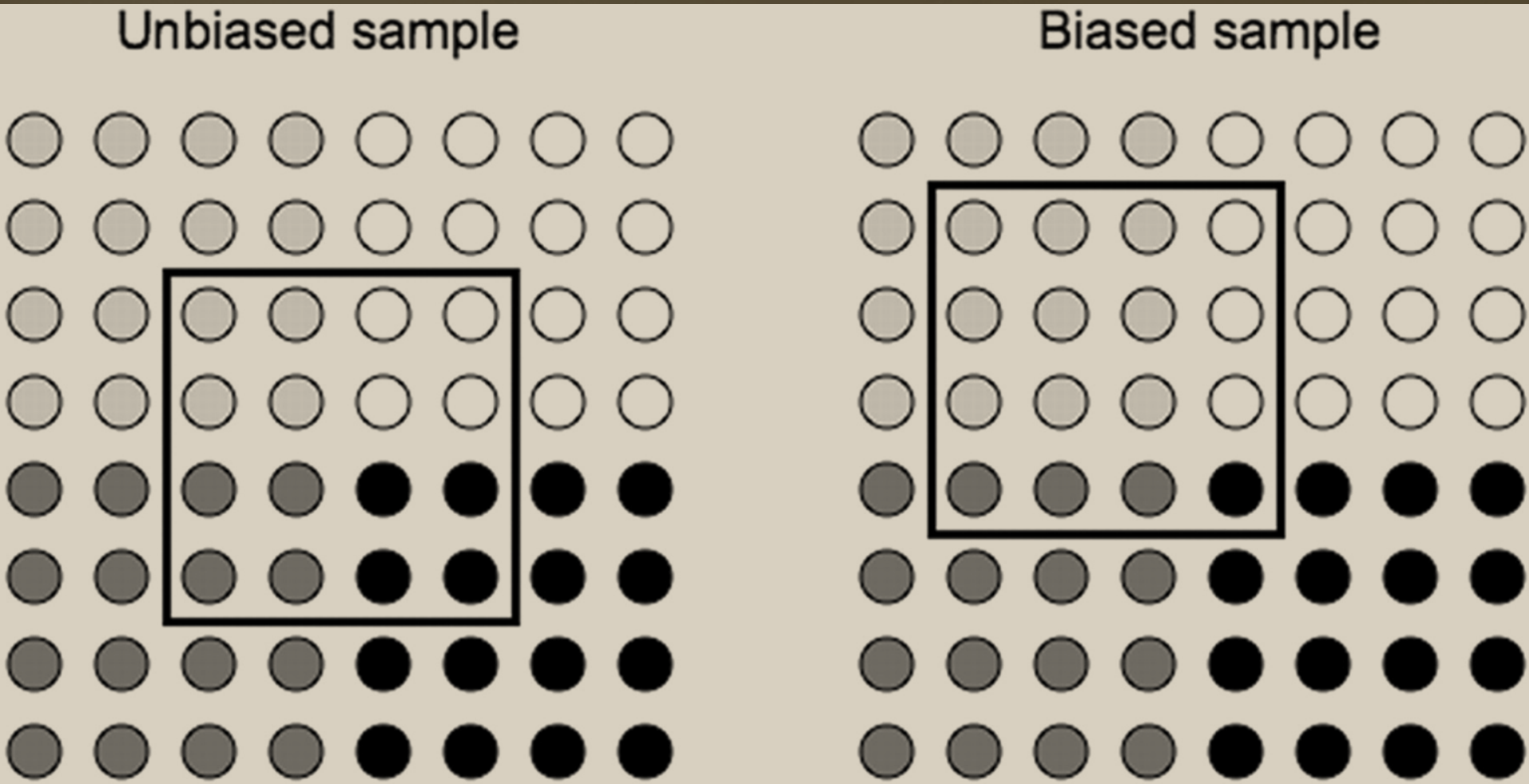
It Could Be that the Purpose of Your Life Is Only to Serve as a Warning to Others.

Observed narrow-but-deep chain letter propagation trees.

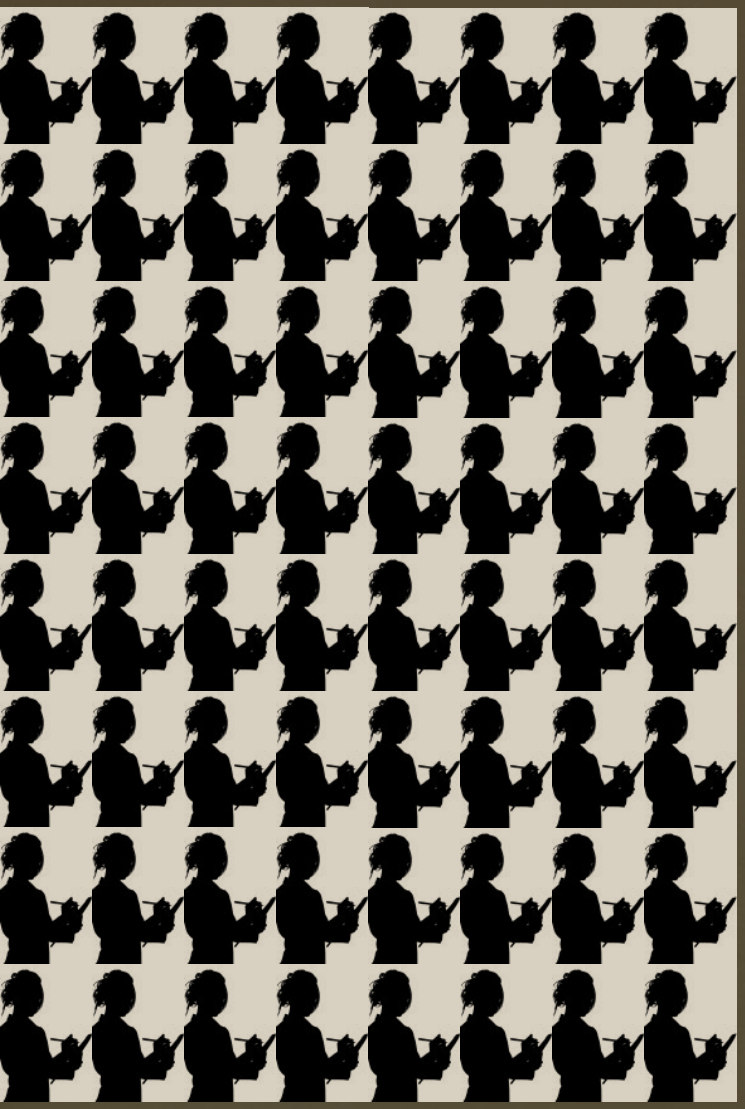
Liben-Nowell, David, and Jon Kleinberg. 2008. "Tracing information flow on a global scale using Internet chain-letter data." *Proceedings of the National Academy of Sciences* 105 (12) (March 25): 4633 -4638. doi:10.1073/pnas.0708471105.

"Selection biases of which data we observe can radically change the estimation of classical diffusion processes."

Golub, B., and M. O. Jackson. 2010. "From the Cover: Using selection bias to explain the observed structure of Internet diffusions." *Proceedings of the National Academy of Sciences* 107 (June 3): 10833-10836. doi:10.1073/pnas.1000814107.



Ravani, Pietro, Patrick S. Parfrey, Bryan Curtis, and Brendan J. Barrett. 2007. "Clinical research of kidney diseases 1: researchable questions and valid answers." *Nephrology Dialysis Transplantation* 22 (12) (December 1): 3681 -3690. doi:10.1093/ndt/gfm838.



Taxpayers

???

Taxpayers
+ Children
+ Others

= Population

Systematic bias in data leads to a systematic bias in conclusions.

If you take a cookie cutter to your research, don't be surprised when a plate of cookies come out of the oven.

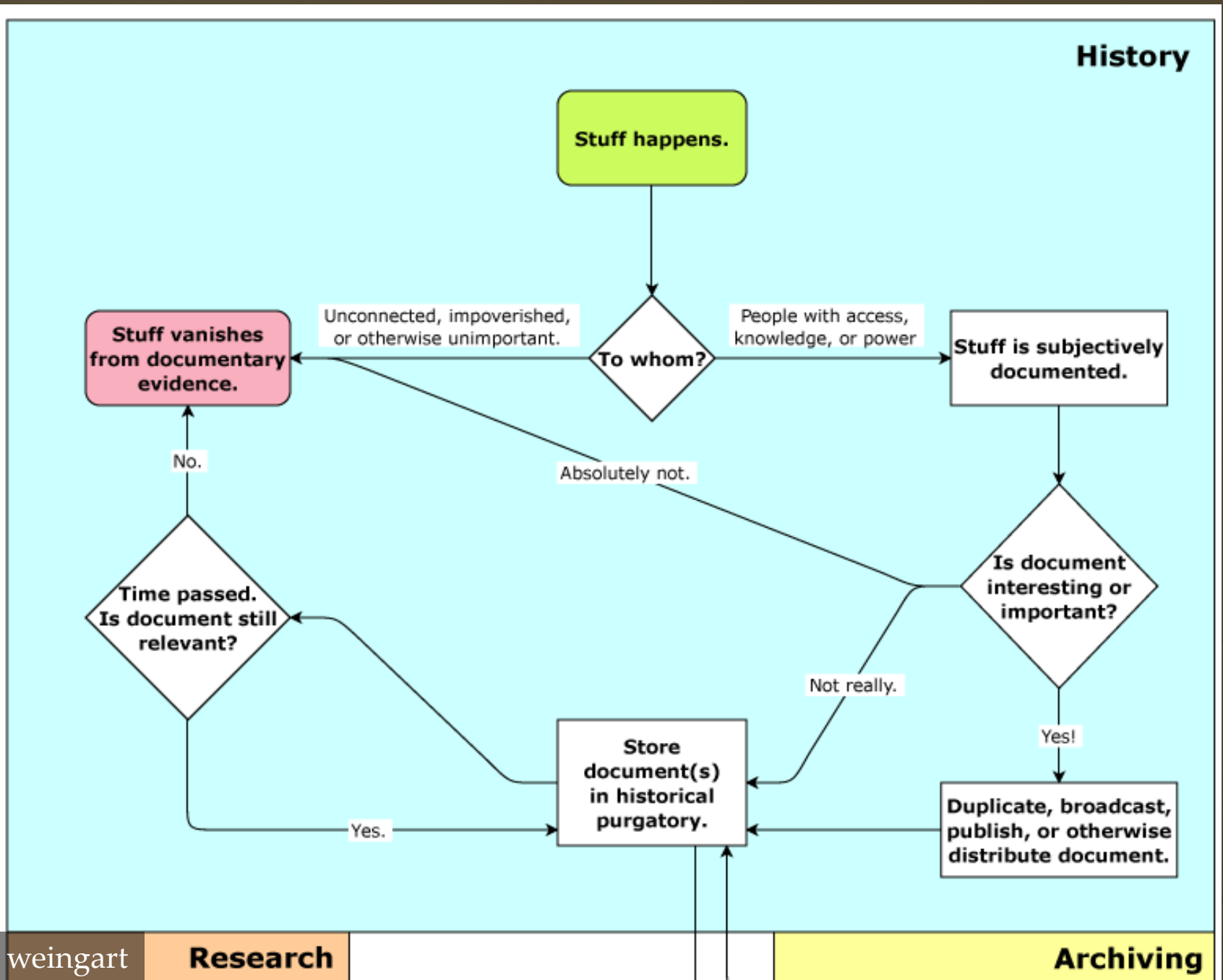
http://www.woohome.com/wp-content/uploads/2008/07/cookie-cutter.jpg

"Once a document has been created, what is the probability that it will survive for *n* years? There is, so far as I know, not a single study which affords an empirical basis for answering this question."

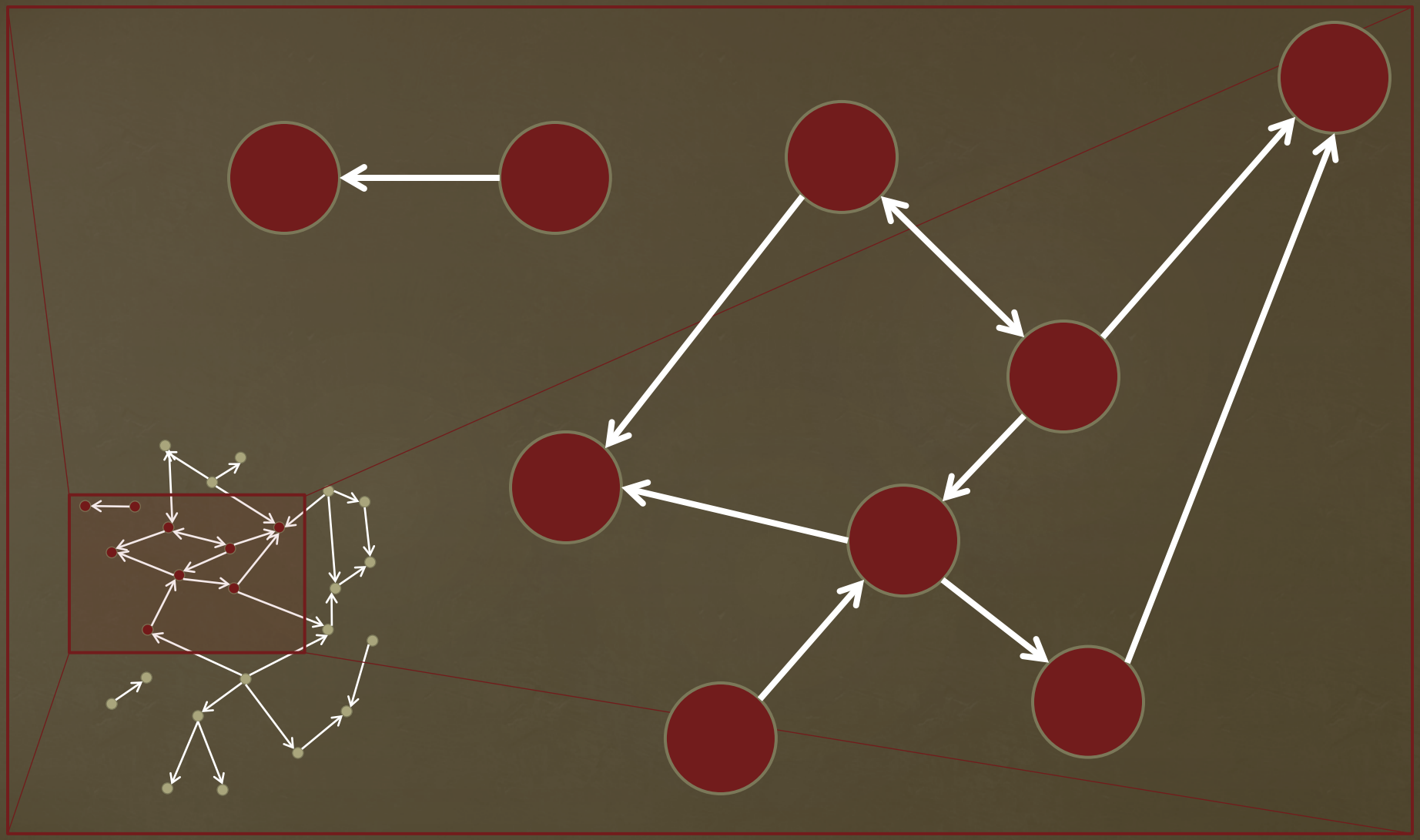
-Murray G. Murphey, 1973

Murphey, Murray G. 1973. *Our Knowledge of the Historical Past*. Macmillan Pub Co.

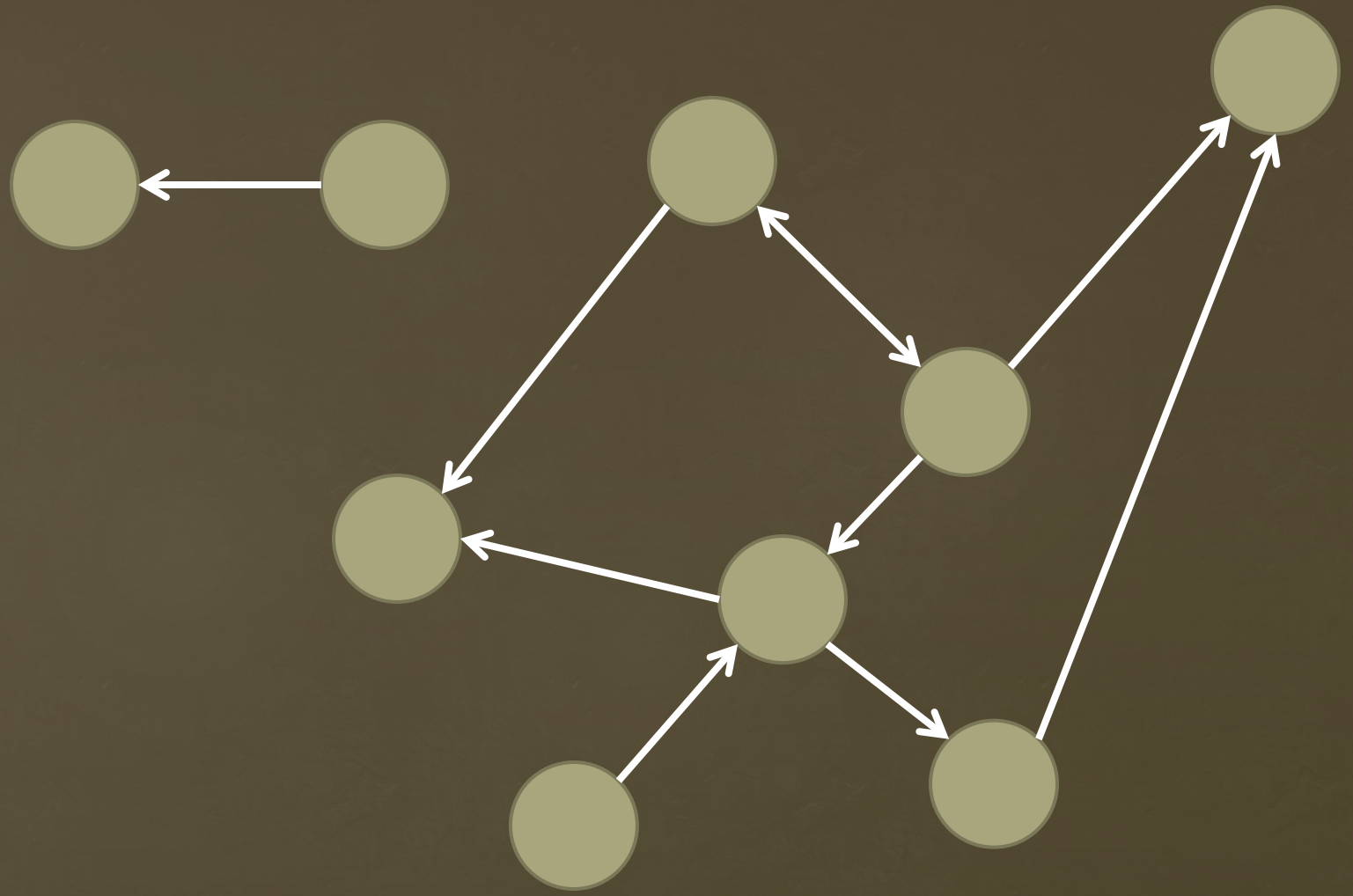
Precise awareness of missing information is *required* for making any large-scale statistical statement.


History
Stuff happens.
To whom?
Unconnected, impoverished, or otherwise unimportant.
People with access, knowledge, or power
Stuff vanishes from documentary evidence.
Stuff is subjectively documented.
No.
Absolutely not.
Is document interesting or important?
Time passed. Is document still relevant?
Not really.
Yes!
Store document(s) in historical purgatory.
Duplicate, broadcast, publish, or otherwise distribute document.
Yes.
Research
Archiving

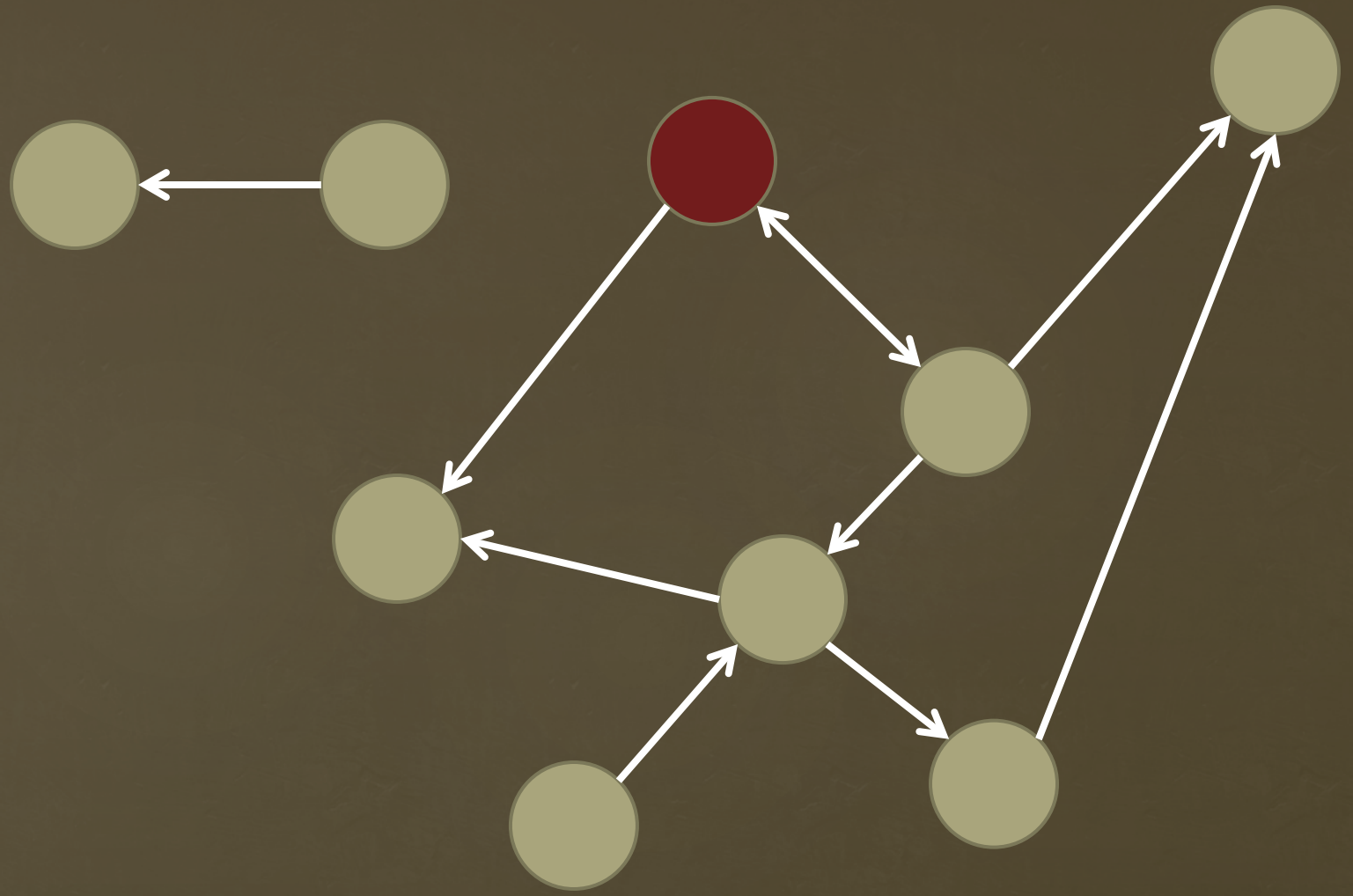
How does biased data affect network analysis?



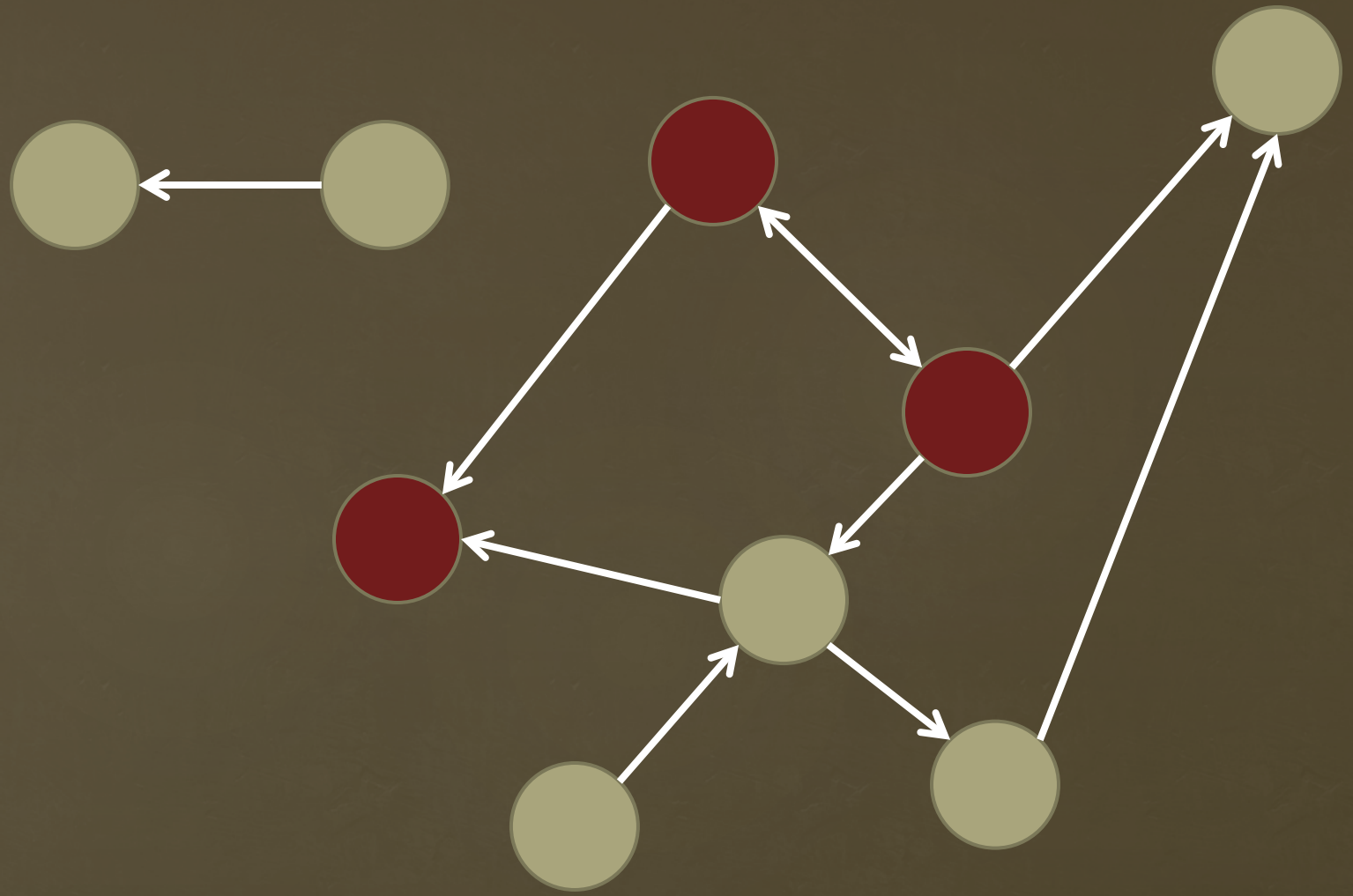
Network Sampling



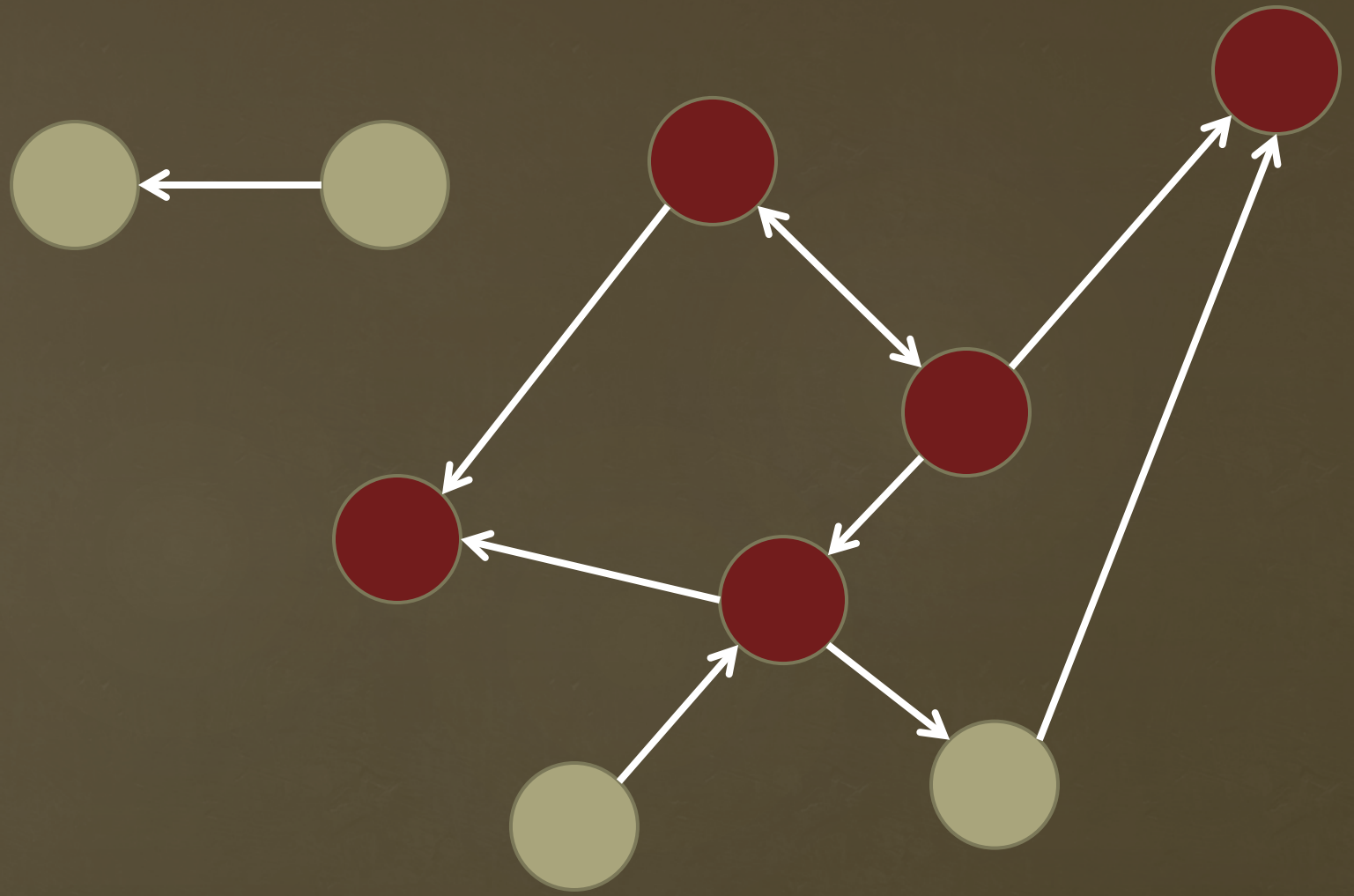
Crawling



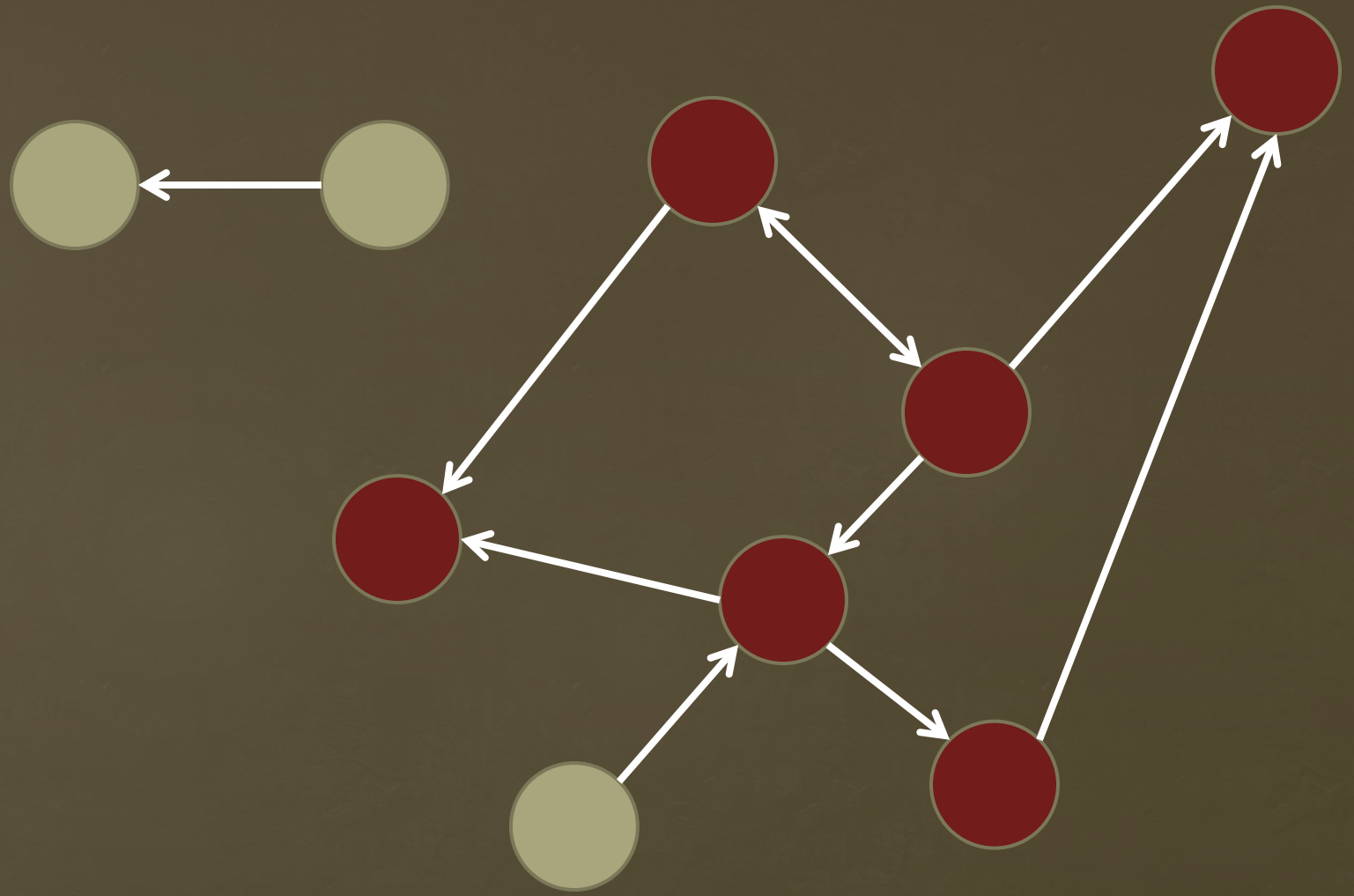
Crawling



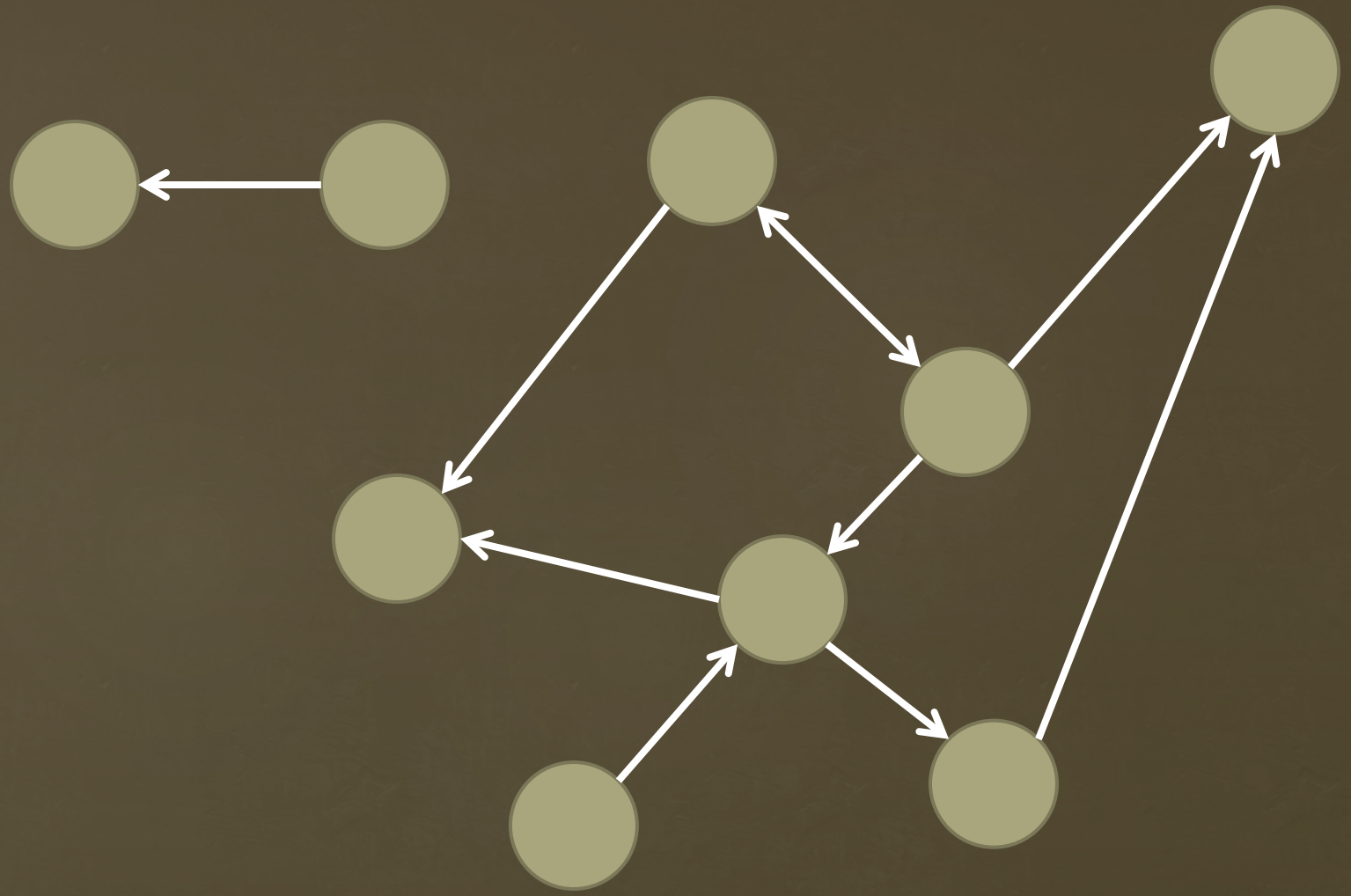
Crawling



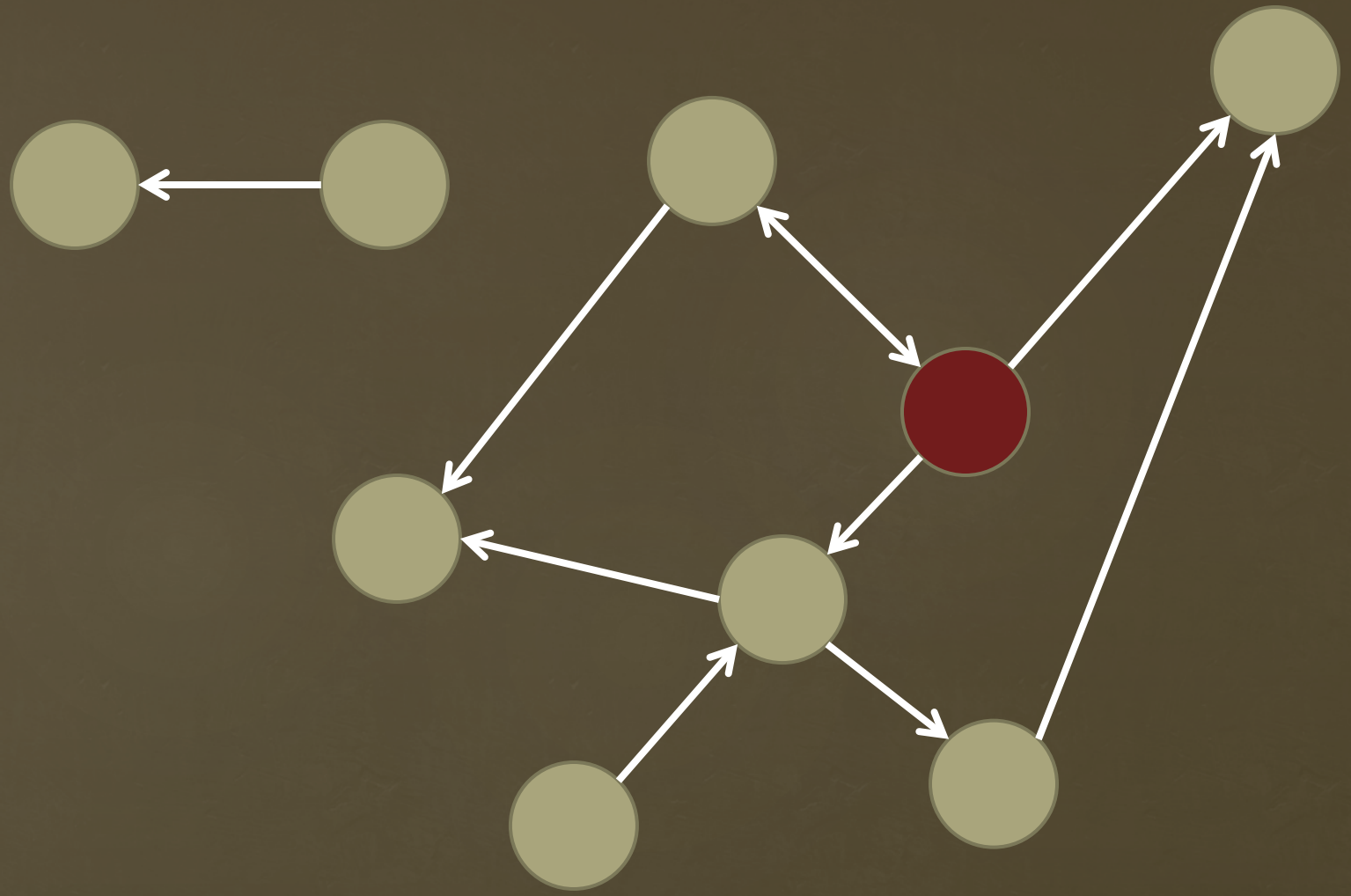
Crawling



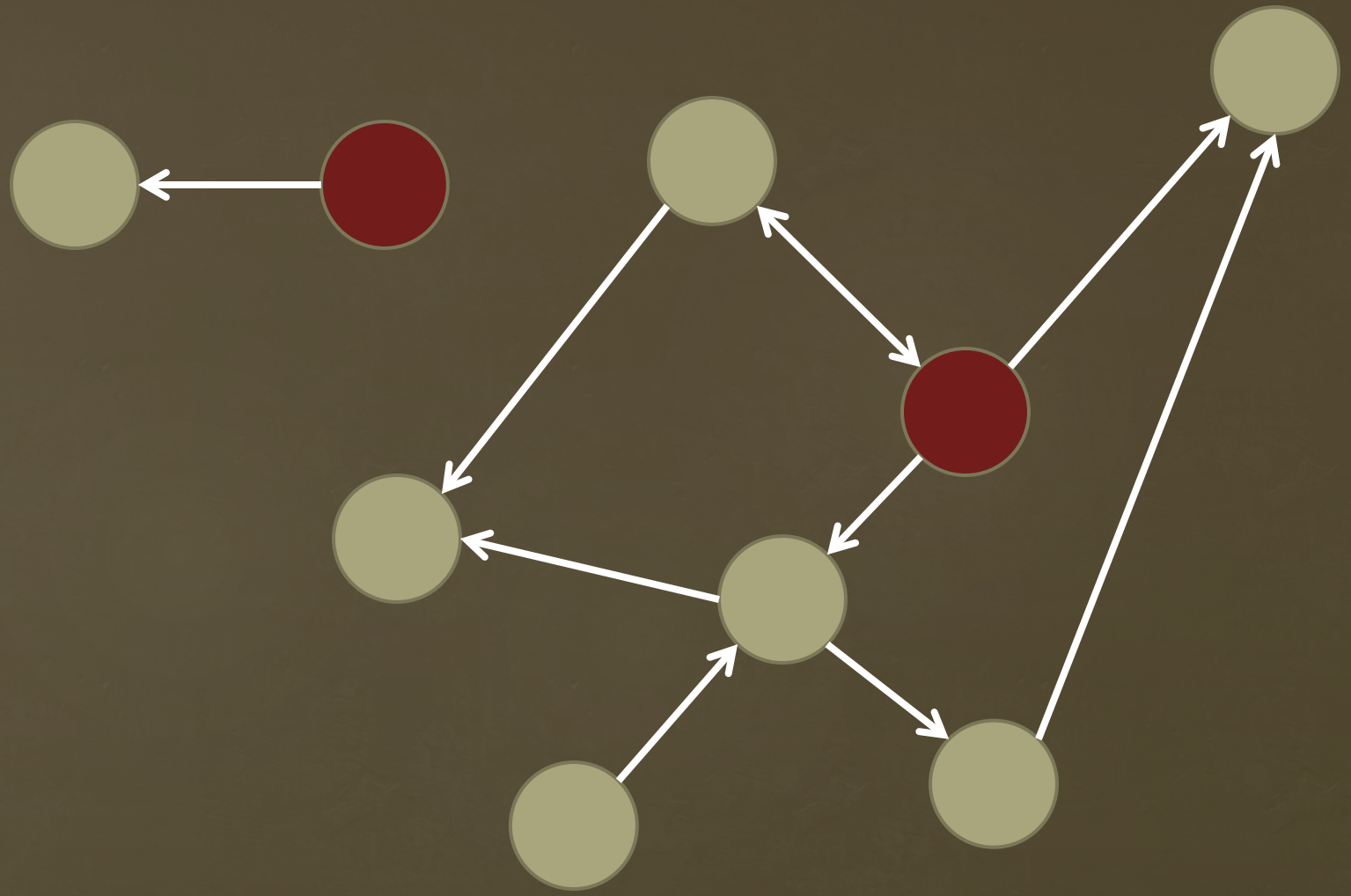
Crawling



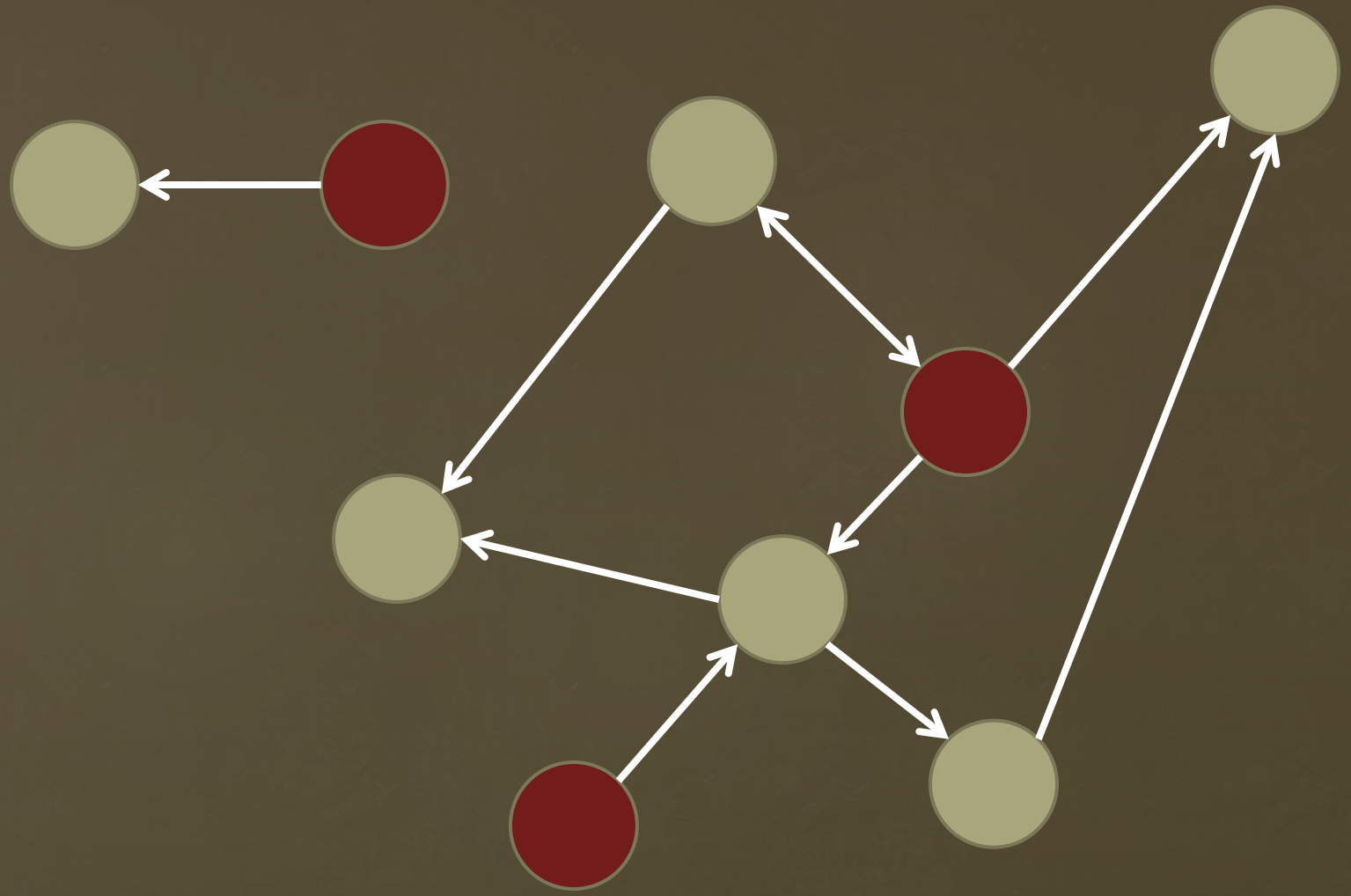
Random Sampling



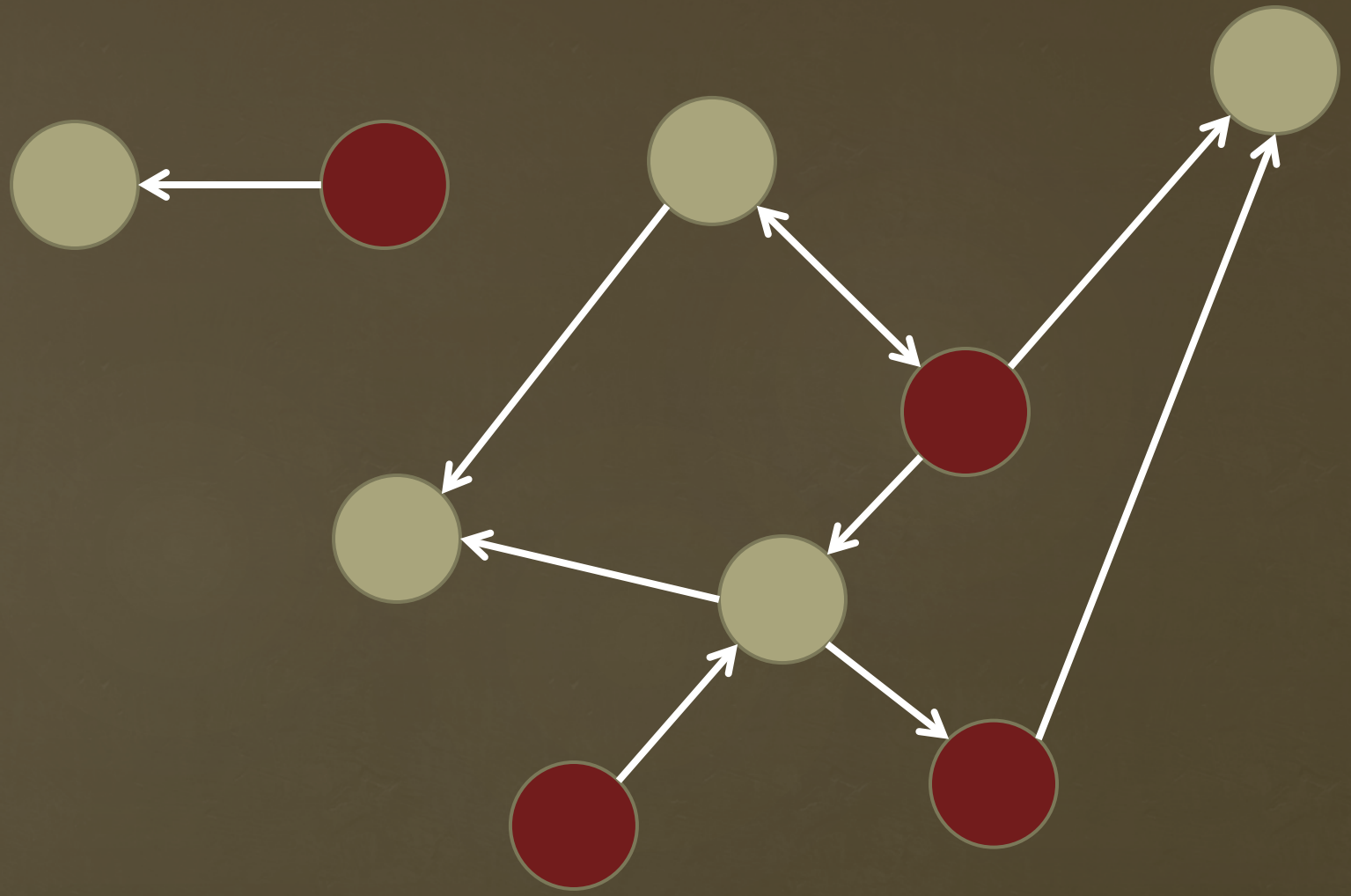
Random Sampling



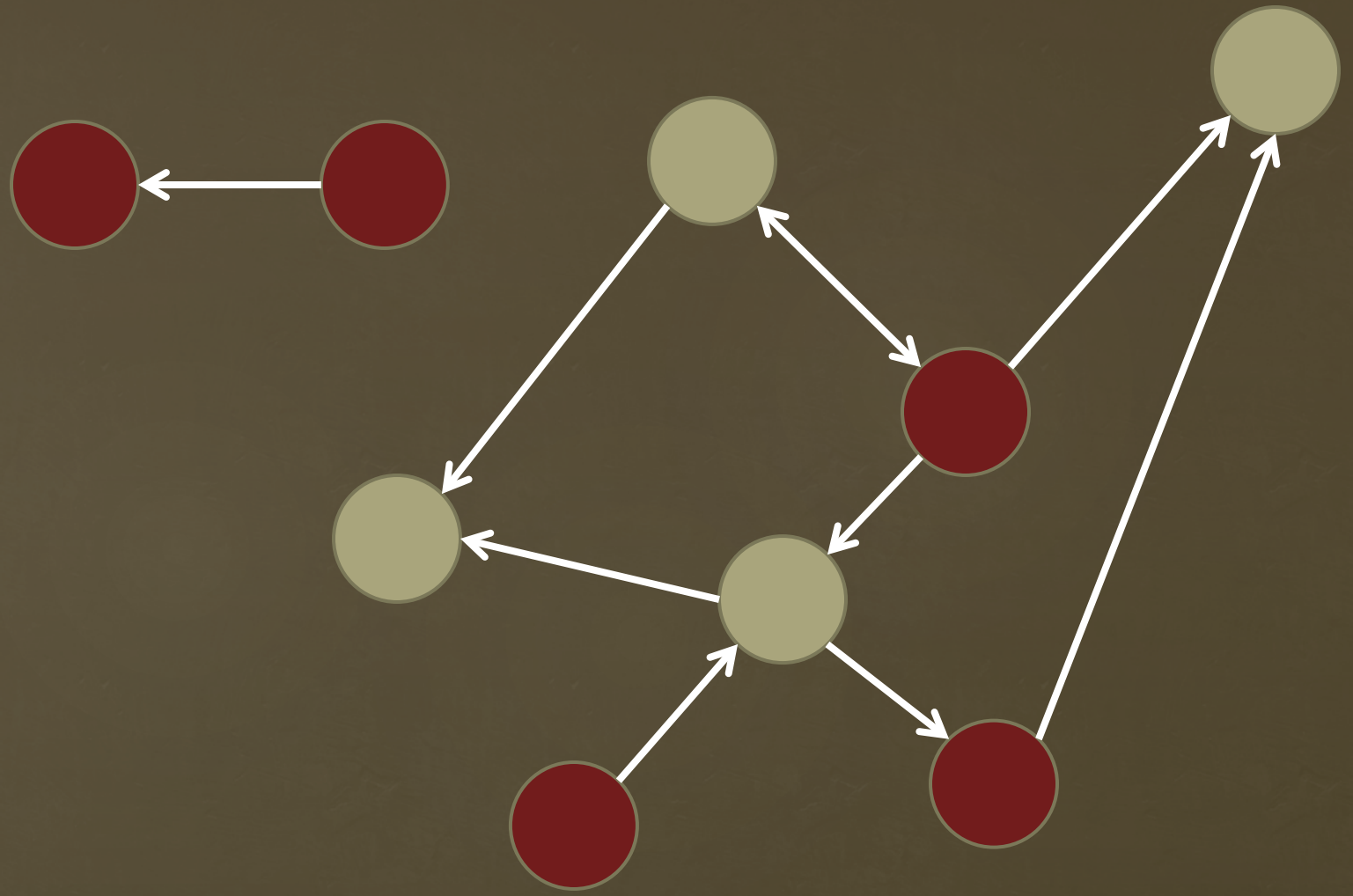
Random Sampling



Random Sampling



Random Sampling



Random Sampling

Crawling Biases

- Nodes with high degree chosen more frequently
- Peripheral nodes on directed networks may never be selected
- Degree-based metrics skewed

Random Sampling Biases

- More disconnected components
- Skewed topology
- Community structure lost

Sailing Ships (c. 1886-1890) Constantinos Volanakis

Sailing Ships (c. 1886-1890) Constantinos Volanakis