Citation Network Analysis

Thom Neale
Twitter: @twneale
Github: twneale
Where to find materials

- The materials for this talk are all located at twneale.github.io/citation-network-analysis
Goodies

- Webpage where you can try out the code I used in this study: https://www.wakari.io/twneale
- App demonstrating (sort of) practical use of network analysis data: https://cite-fight.com
Big Picture

- How to get from unstructured text to data that people use to create excellent tools
This talk is about

- The journey from unstructured text to highly structured data
- Fantasizing about the amazing things we can do once the journey is complete
Consider the following question:
What determines the importance of a case?
Michael Gerhardzt observed that the extent and nature of a precedent’s network of citations determine the strength of its constraining power on subsequent cases. He argued further that the authority of a precedent depends on the consistency and uniformity with which other authorities have cited it.

Network of citations?

Proceedings diagram:

- Planned Parenthood of Southeastern PA v. Casey, 505 U.S. 833 (1992)
- Roe v. Wade, 410 U.S. 113 (1973)

Fowler et al., *The Authority of Supreme Court Precedent* (2008)
Translation:

- Constraining power of a case on subsequent cases (i.e., importance) depends on two things:
  1) Nature and extent of case's network of citations
  2) Consistency and uniformity with which other authorities have cited it
1) The extent and nature of a precedent’s network of citations

- Webster v. Reproductive Health Services
  492 U.S. 490 (1989)

- Akron v. Akron Center for Reproductive Health, Inc.
  462 U.S. 416 (1983)

- Planned Parenthood of Southeastern PA. v. Casey
  505 U.S. 833 (1992)

- Roe v. Wade
  410 U.S. 113 (1973)

- Thornburgh v. American College of Obstetricians and Gynecologists
  476 U.S. 747 (1986)
Conclusion

Planned Parenthood v Casey is “well founded in law”
2) Consistency and uniformity with which other authorities have cited it.

- Planned Parenthood of Southeastern PA v. Casey, 505 U.S. 833 (1992)
- Roe v. Wade, 410 U.S. 113 (1973)
Conclusion:

*Roe v Wade* is “influential”
Turn that into an algorithm
Make it recursive...
Jon M. Kleinberg

Authoritative Sources in a Hyperlinked Environment (1998)


“Hyperlink-induced Topic Search (HITS)
PageRank is another
Degree Centrality is Another
Recap
Case Citations form a Network

Webster v. Reproductive Health Services
492 U.S. 490 (1989)

Akron v. Akron Center for Reproductive Health, Inc.
462 U.S. 416 (1983)

Planned Parenthood of Southeastern PA. v. Casey
505 U.S. 833 (1992)

Roe v. Wade
410 U.S. 113 (1973)

Thornburgh v. American College of Obstetricians and Gynecologists
476 U.S. 747 (1986)
We can use network analysis algorithms to rank the nodes in the network.
Numerous Algorithms Exist

- Indegree Centrality
- PageRank
- HITS (Hyperlink-Induced Topic Search)
- Eigenvector Centrality
- And many, many, other stupefying algorithms
The network of case citations is "scale-free"
Random network
Other important scale-free networks?
THE INTERNET
Network rankings change over time
Time-series data creates interesting opportunities

Methodology

The charts above show the yearly indegree centrality scores of each case from the time it was published to the present.

The data for this app was obtained from CanLII's new API.

The trend value for each case was calculated by fitting a linear polynomial function to the yearly indegree centrality scores.
You can fit curves to the data
The curves enable you to estimate things.
What things?

- Is this case's importance increasing or decreasing? (slope, derivative)
- Which of these cases has had a greater cumulative influence over time? (area)
- Does anyone still use this case? (x intercept)
Is this case's importance increasing or decreasing?

\[ f(x) = x \cdot \sin(x^2) + 1 \]
\[ f'(x) = \sin(x^2) + 2x^2 \cdot \cos(x^2) \]

Points:
- \((0, 0)\)
- \((1.3552, 2.3076)\)
- \((2.1945, -1.1828)\)
- \((2.8137, 3.8081)\)
Which of these cases has had a greater cumulative influence over time?
Does anyone still use this case?
Findings

- On average, Canada Supreme Court (CSC) cases “fail” after 50 years.
- About 18% of CSC cases have survived longer than 15 years (and is still positive?).
- In all other courts, the average time to failure ranges from 3 to 15 years.
- In all other courts, less than 3% of cases survive longer than 15 years.
Challenges

- Citation Extraction is hard
- Resolving citations to sources is harder
Citation Extraction is hard

- Regular expressions are quick and easy
- But they don't scale
- Regexes alone aren't good at processing highly variable patterns
- “citation parsing” is a special case of entire document parsing
- Nested data structures (like citations) require stateful, recursive code
Regexes are neither stateful nor recursive
“There is some case law suggesting (without much discussion) that a purchaser cannot maintain a caveat unless it can be shown that specific performance is available. Where there is no binding contract, such that the purchaser is unable to get any remedy, clearly a caveat cannot be maintained: Oxford Development Group Inc. v. Midland Development Ltd., [1993] A.J. No. 47 (C.A.).”
“There is some case law suggesting (without much discussion) that a purchaser cannot maintain a caveat unless it can be shown that specific performance is available. Where there is no binding contract, such that the purchaser is unable to get any remedy, clearly a caveat cannot be maintained: Oxford Development Group Inc. v. Midland Development Ltd., [1993] A.J. No. 47 (C.A.).”
“There is some case law suggesting (without much discussion) that a purchaser cannot maintain a caveat unless it can be shown that specific performance is available. Where there is no binding contract, such that the purchaser is unable to get any remedy, clearly a caveat cannot be maintained: *Oxford Development Group Inc. v. Midland Development Ltd.*, [1993] A.J. No. 47 (C.A.).”
Becomes this
Then This

-Node([])
  -Content([[(0, Token.Content, u'There is some case law suggesting (without much discussion) that a purchaser cannot maintain a caveat unless it can be shown that specific performance is available. Where there is no binding contract, such that the purchaser is unable to get any remedy, clearly a caveat cannot be maintained: ')]])
  -Source([])
  -Title([[(297, Token.Title, u'Oxford Development Group Inc. v. Midland Development Ltd.')]])
  -Citations([])
  -Citation([])
    -SlipYear([[(356, Token.SlipYear, u'[1993]')]])
    -Reporter([[(363, Token.Reporter, u'A.J. No.')]])
    -SlipNumber([[(372, Token.SlipNumber, u'47')]])
    -Jurisdiction([[(375, Token.ParenAbbrev, u'(C.A.)')]])
  -Content([[(381, Token.Content, u'; ')]])
Hey wait! That's a network!
Citation Data Should Probably Be Stored as a Graph

- Complex queries will be cheap, rather than impossible
- More information gets stored, probably in less space
Really Hard

- Resolving citations back to sources
- Why:
  - Volume pages are not unique identifiers
  - Titles aren't unique identifiers either
  - An alarming percentage of citations contain typos
Volume Pages aren't unique IDs
Comparing Titles

- Is unreliable
- Even hard for

At Minimum

- To resolve book citations to sources, you need:
  - Detailed metadata about the book volumes
  - Need it PER VOLUME! Yes, it can change from volume to volume
  - Are cases tabular? Full text? Discretely paginated? Continuously paginated?
Need Publication Metadata
It might look like this
Wait a minute...
It might look like this
None of this is useful. Take a yellow pad back to your office and create some value for once.
An Overcomplicated Hypo
Instead of comparing titles
Ask the neighbors
Recap

- We discussed a strategy for resolving ambiguous book citations that didn't require title comparison.

- It was only possible because our graph database contains 1) publication metadata (report, series, volume, pagination), 2) cases, structured as subgraphs.
Not So Fast

- A horrifying percentage of citations contain typos, and resolving them to a source because even more ridiculous
- We have to try to reverse engineer the typo, then repeat the process for each candidate
- Title comparison would probably be helpful here too
The Technology

- Python
- Networkx
- EC2
- Celery
- Numpy, SciPy
- neo4j
Further Reading

- Programming the Semantic Web (Toby Segaran)
- Collective Intelligence (Toby Segaran)
- Graph Databases (Ian Robinson)
- Machine Learning for Hackers (Drew Conway)
The End