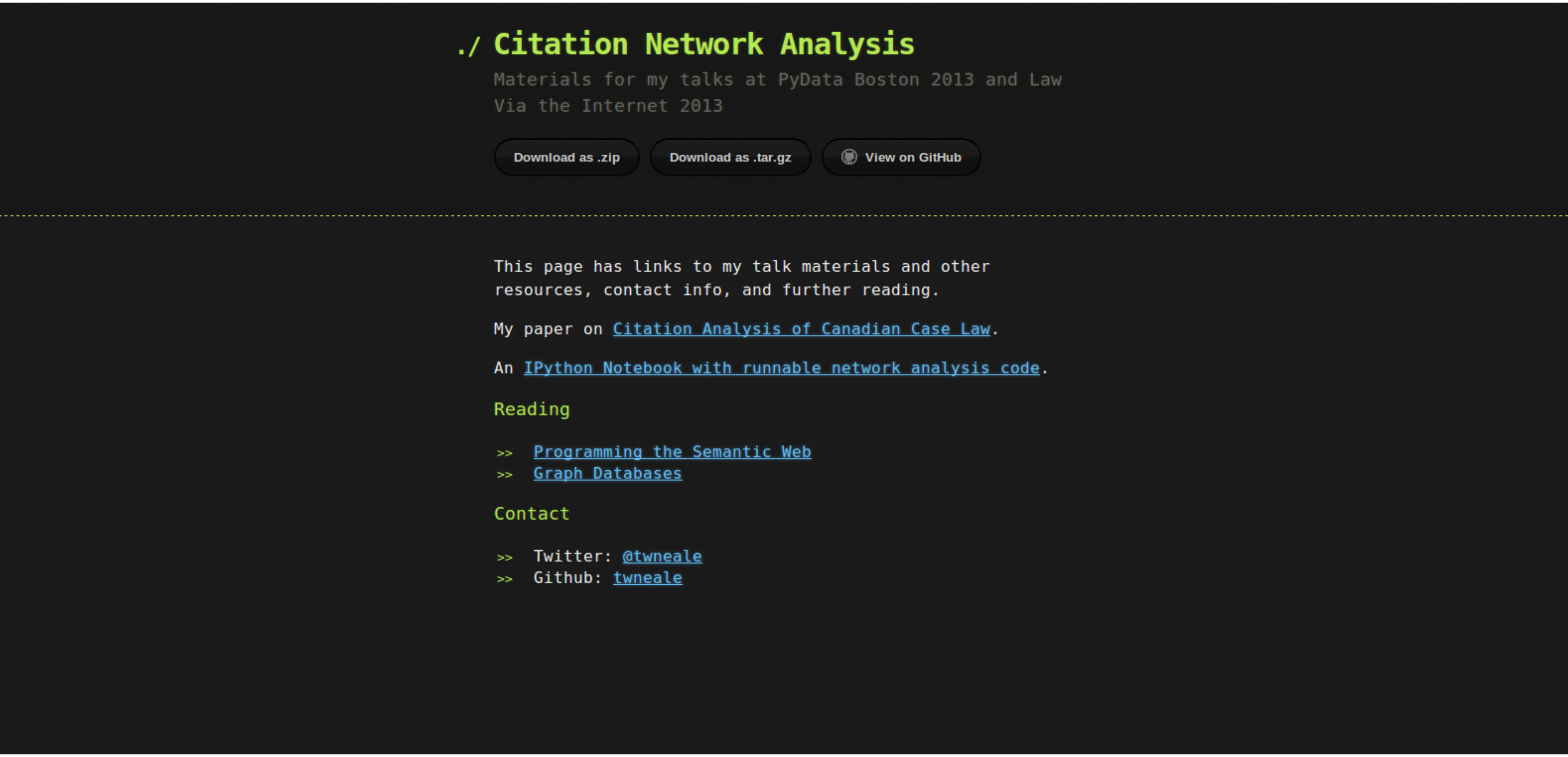
# Citation Network Analysis

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



## ./ Citation Network Analysis

Materials for my talks at PyData Boston 2013 and Law Via the Internet 2013

This page has links to my talk materials and other resources, contact info, and further reading.

My paper on Citation Analysis of Canadian Case Law.

An IPython Notebook with runnable network analysis code.

### Reading

>> Programming the Semantic Web
>> Graph Databases

### Contact

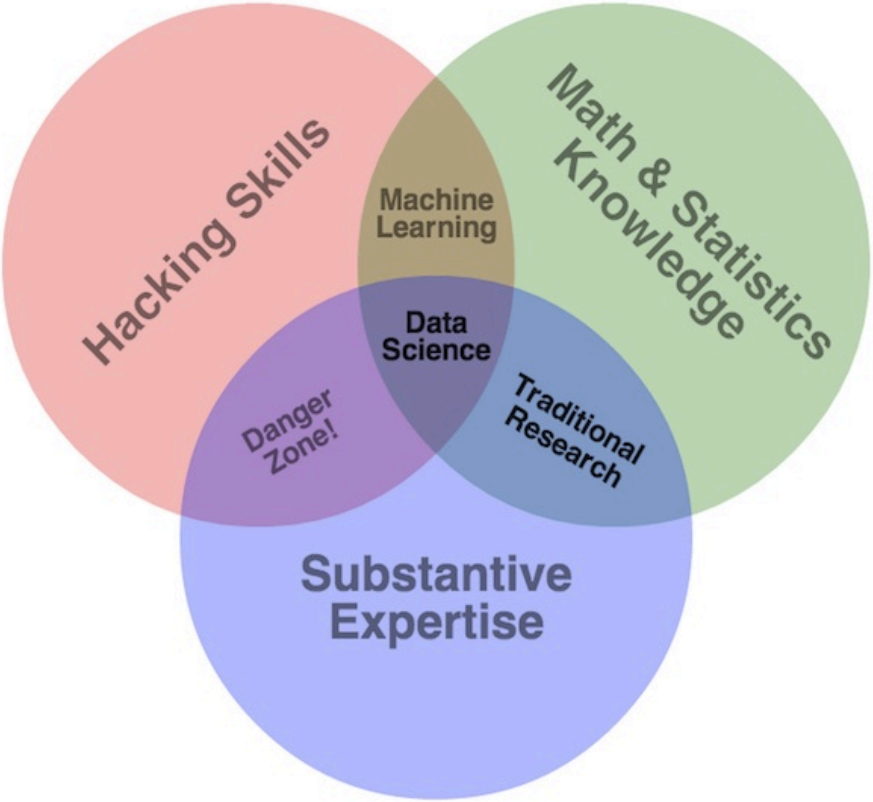
>> Twitter: @twneale
>> Github: twneale

# 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



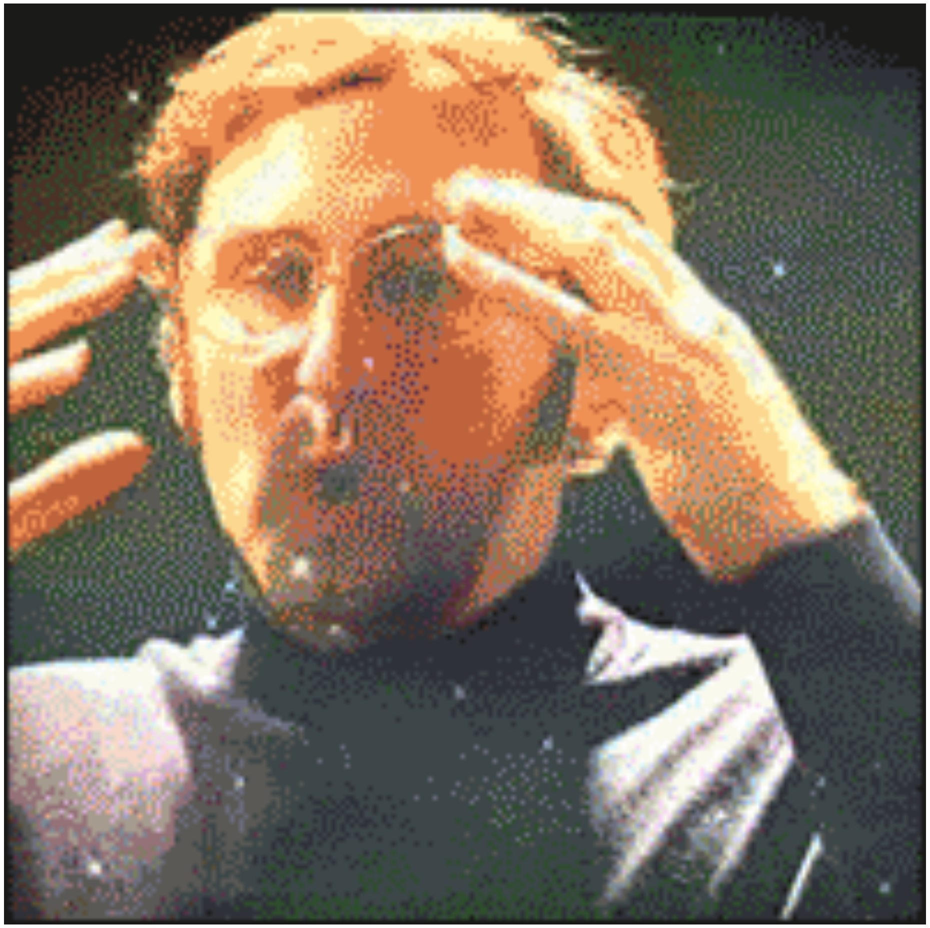
drewconway.com

# 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?

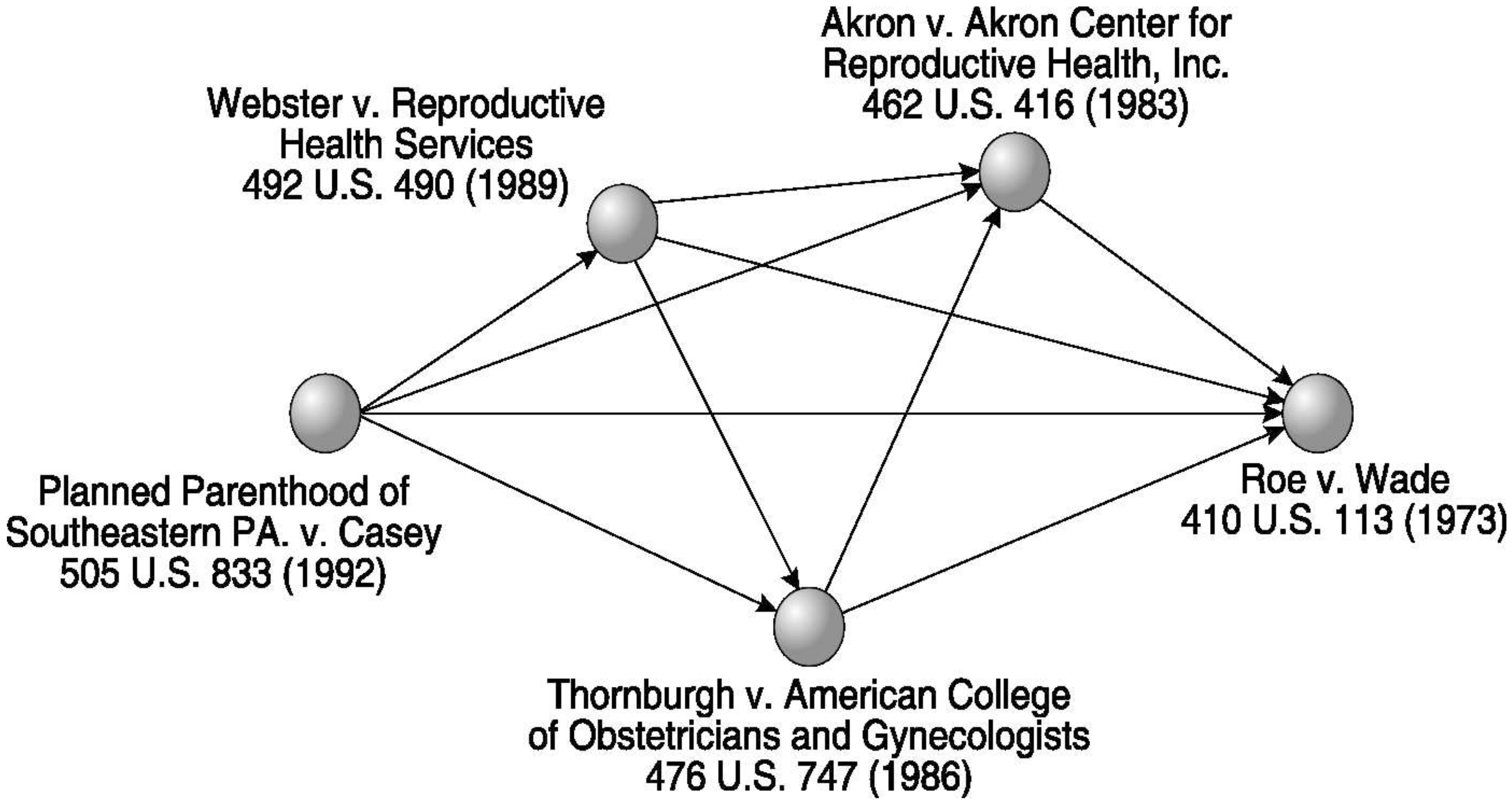


# Maybe this:

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.

Michael J. Gerhardt, The Irrepressibility of Precedent, 86 N.C.L.REV.1279, 1291 (2008)

# Network of citations?



Fowler et at, *The Authority of Supreme Court Precedent* (2008)

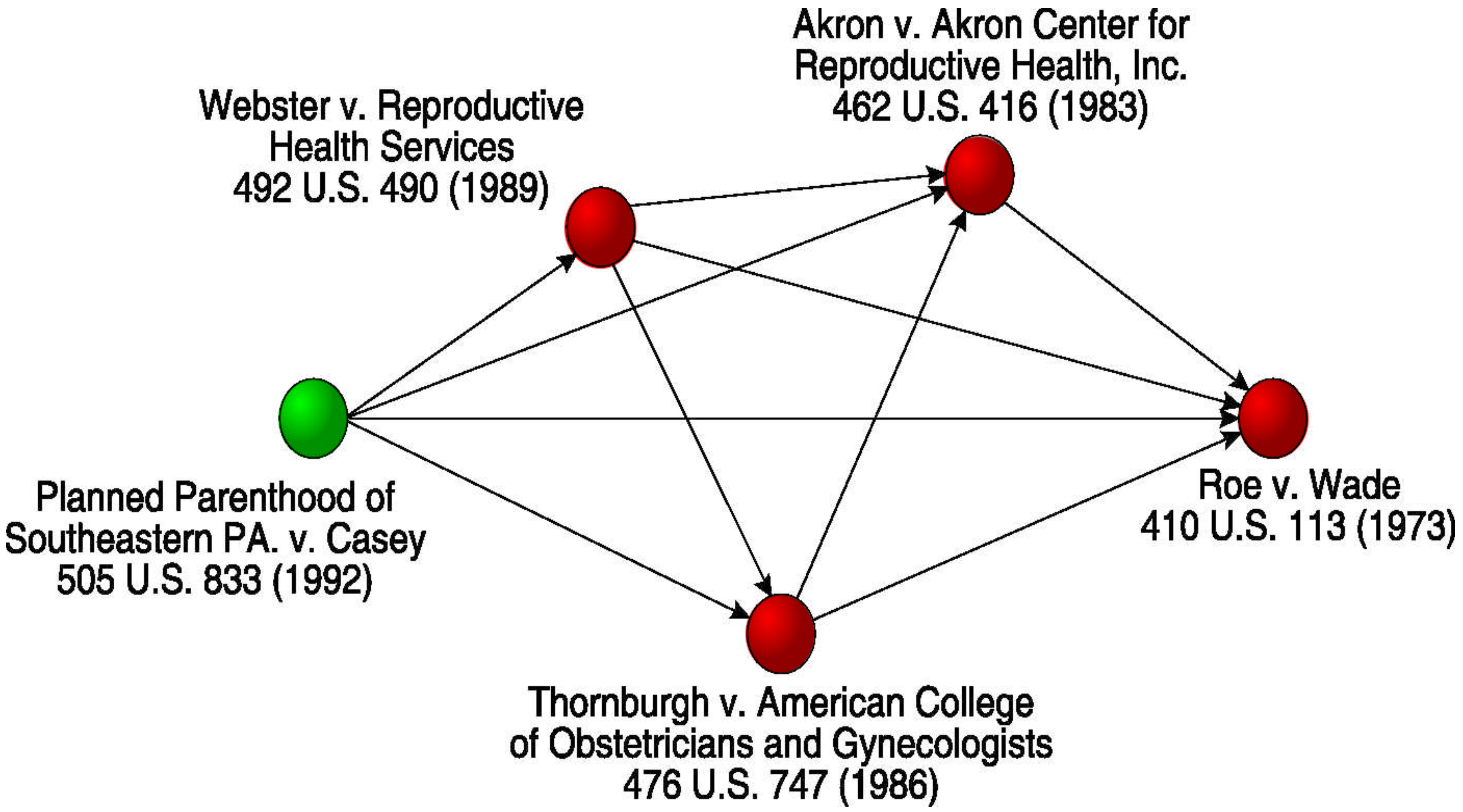
# Translation:

- 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

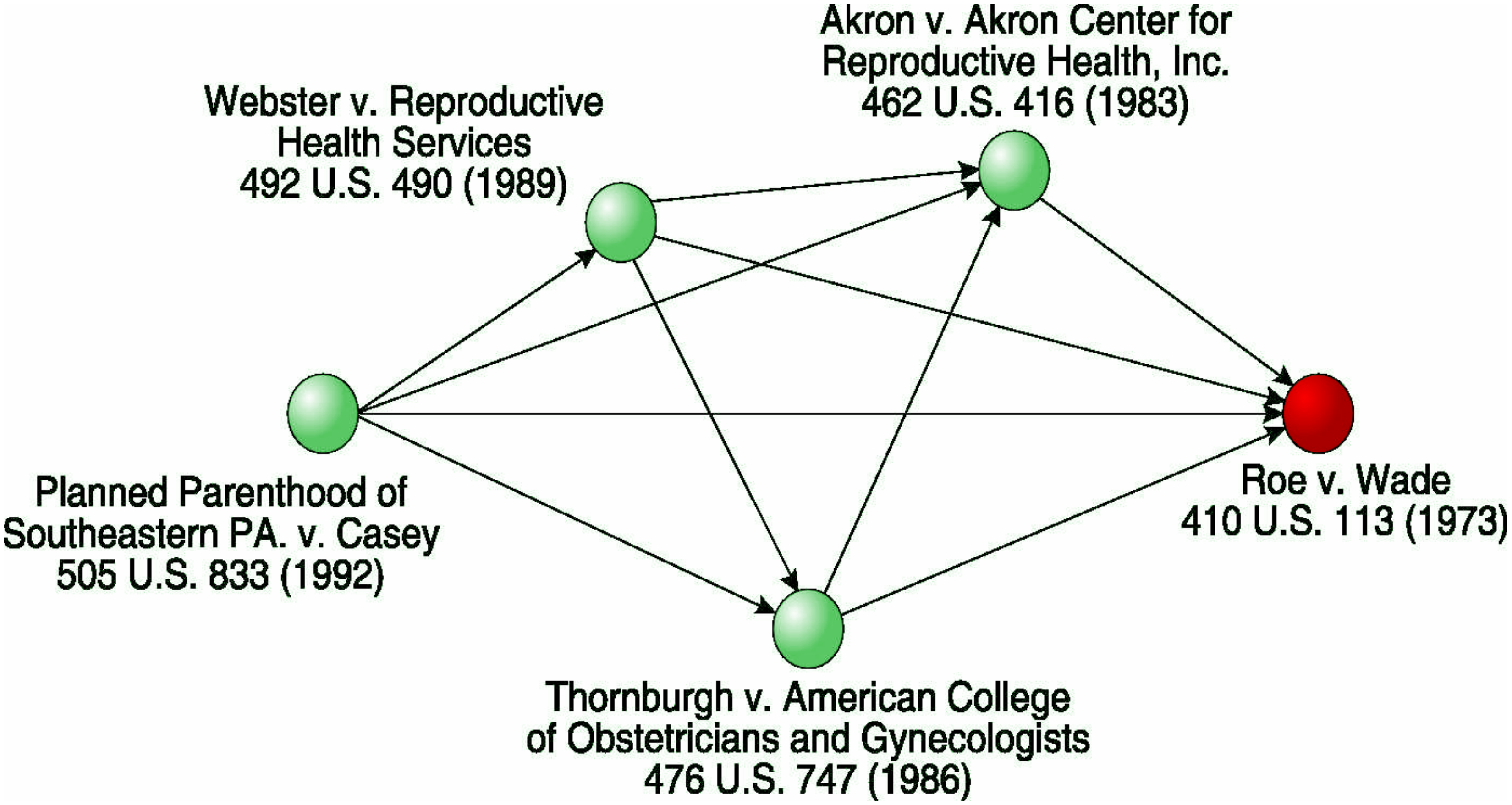


# Conclusion

*Planned Parenthood v Casey* is

“well founded in law”

# 2) Consistency and uniformity with which other authorities have cited it.

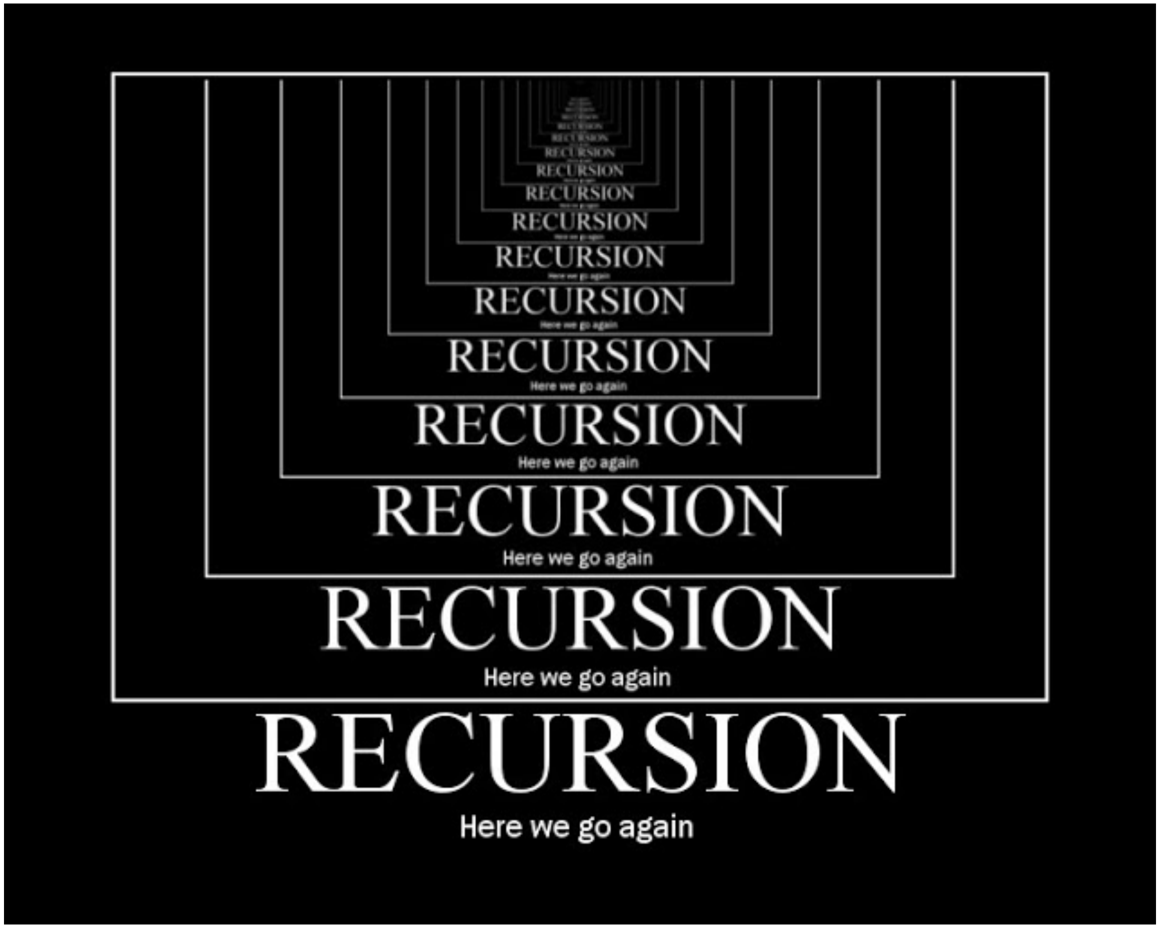

Akron v. Akron Center for Reproductive Health, Inc. 462 U.S. 416 (1983)
Webster v. Reproductive Health Services 492 U.S. 490 (1989)
Planned Parenthood of Southeastern PA. v. Casey 505 U.S. 833 (1992)
Roe v. Wade 410 U.S. 113 (1973)
Thornburgh v. American College of Obstetricians and Gynecologists 476 U.S. 747 (1986)

# Conclusion:

*Roe v Wade* is “influential”

# Turn that into an algorithm

# Make it recursive...



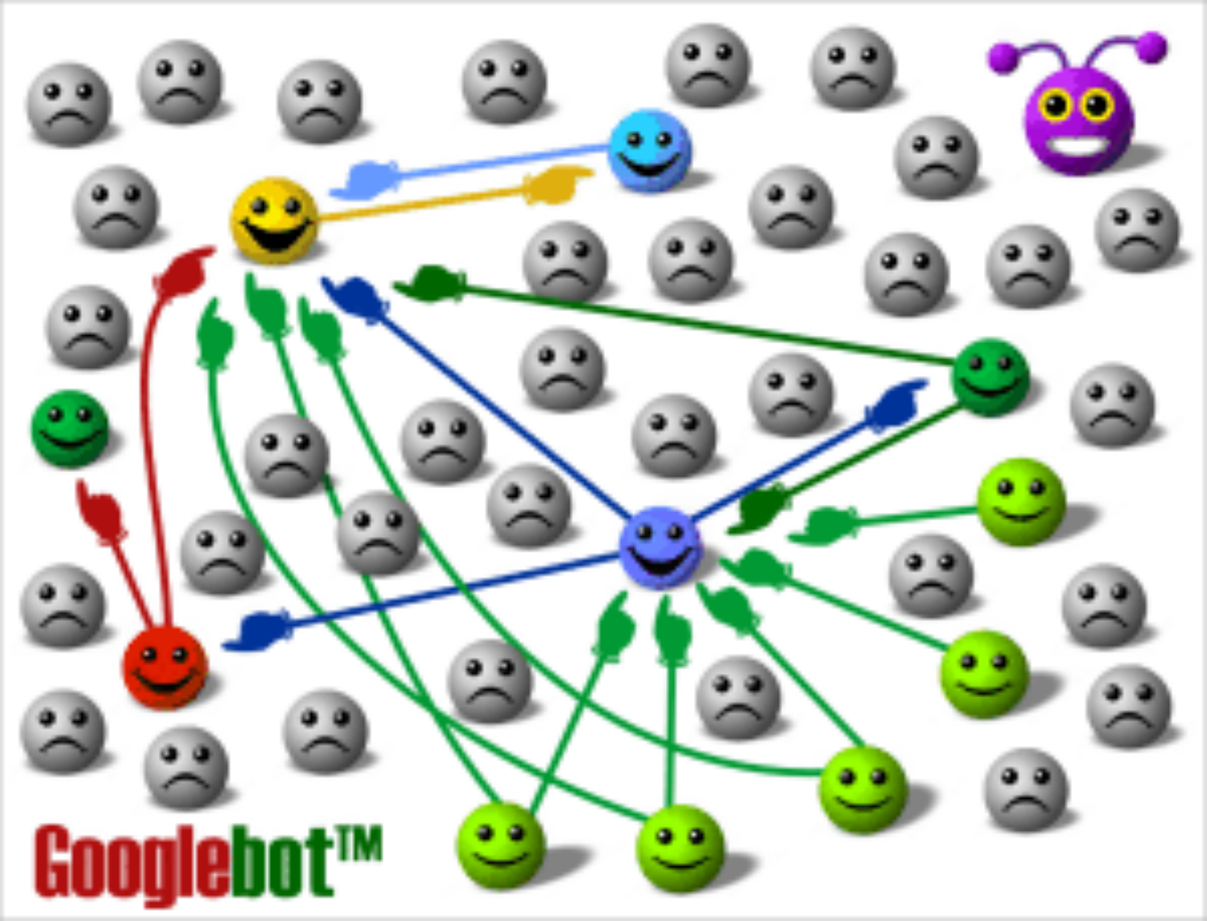
# Jon M. Kleinberg

Authoritative Sources in a Hyperlinked Environment (1998)

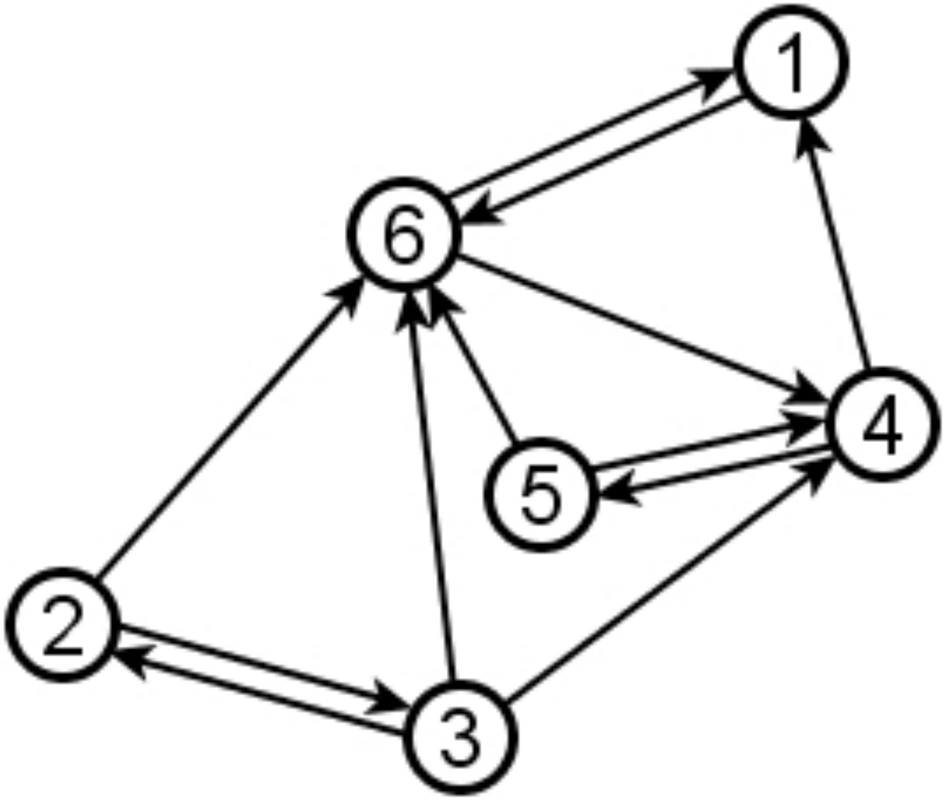
http://www.cs.cornell.edu/home/kleinber/auth.pdf

“Hyperlink-induced Topic Search (HITS)

# PageRank is another

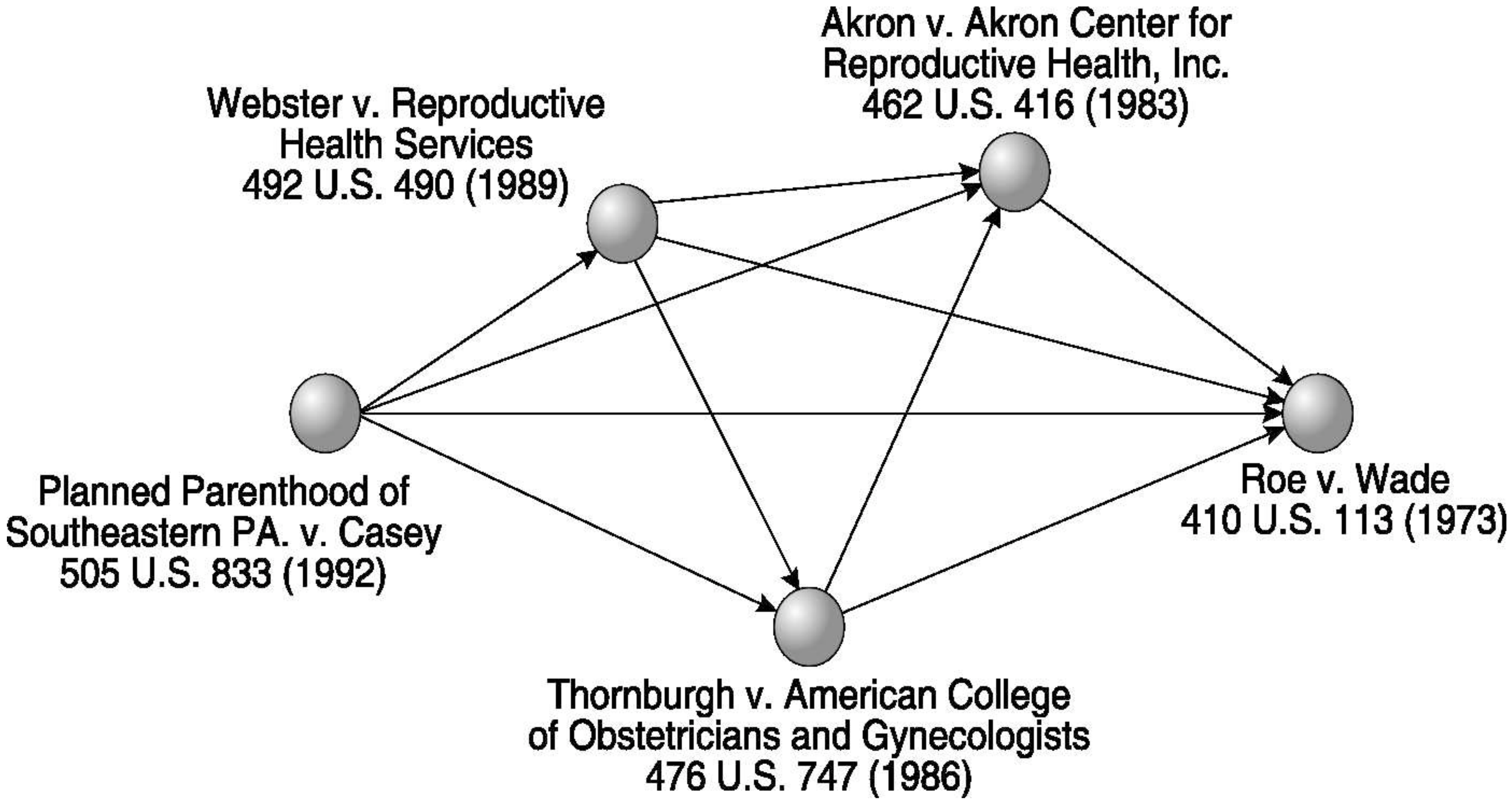


# Degree Centrality is Another

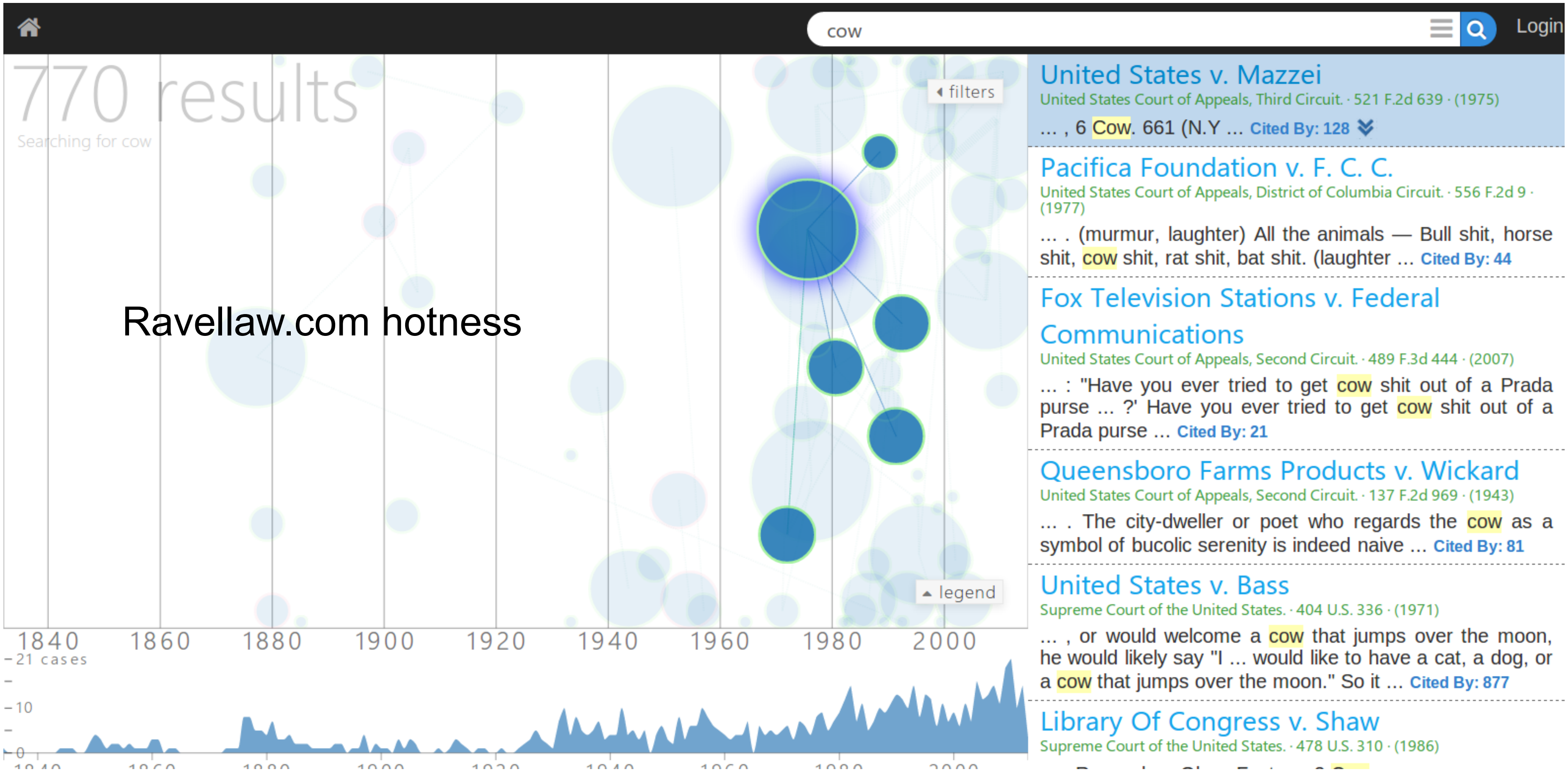


# Recap

# Case Citations form a Network



# We can use network analysis algorithms to rank the nodes in the network



Ravellaw.com hotness

**United States v. Mazzei**
United States Court of Appeals, Third Circuit. · 521 F.2d 639 · (1975)
... , 6 Cow. 661 (N.Y ... Cited By: 128

**Pacifica Foundation v. F. C. C.**
United States Court of Appeals, District of Columbia Circuit. · 556 F.2d 9 · (1977)
... . (murmur, laughter) All the animals — Bull shit, horse shit, cow shit, rat shit, bat shit. (laughter ... Cited By: 44

**Fox Television Stations v. Federal Communications**
United States Court of Appeals, Second Circuit. · 489 F.3d 444 · (2007)
... : "Have you ever tried to get cow shit out of a Prada purse ... ?' Have you ever tried to get cow shit out of a Prada purse ... Cited By: 21

**Queensboro Farms Products v. Wickard**
United States Court of Appeals, Second Circuit. · 137 F.2d 969 · (1943)
... . The city-dweller or poet who regards the cow as a symbol of bucolic serenity is indeed naive ... Cited By: 81

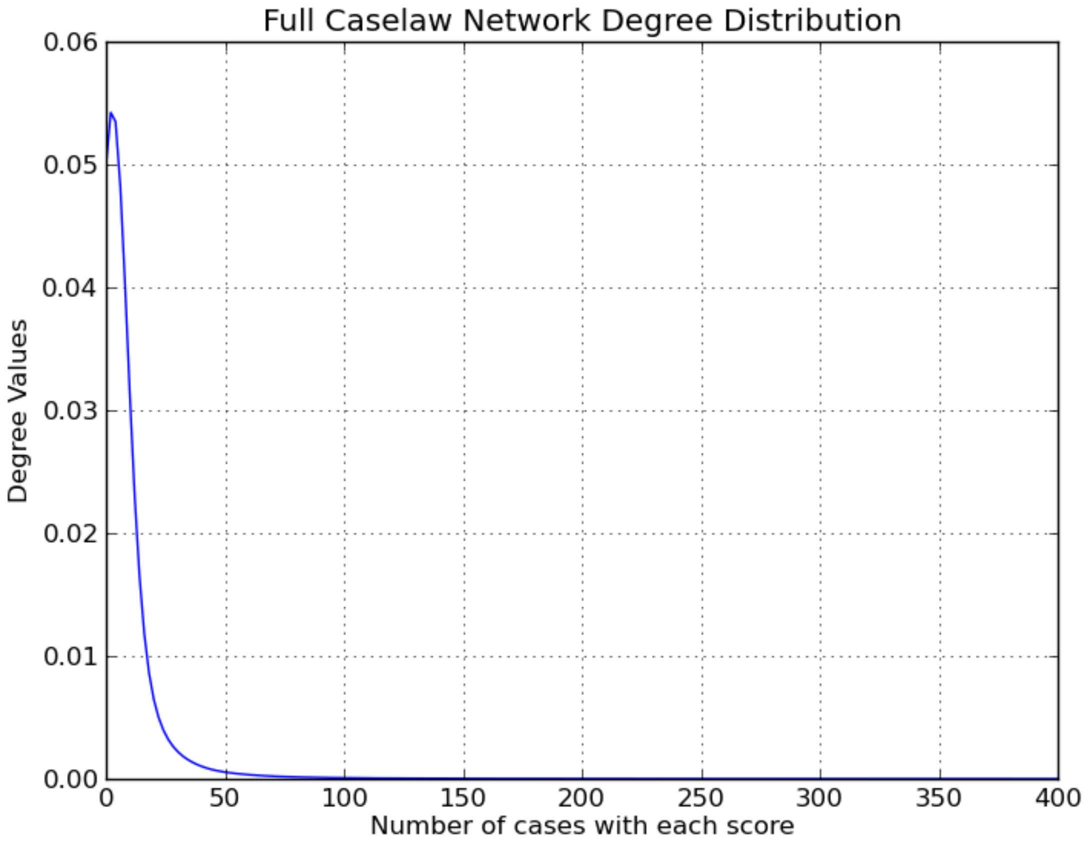
**United States v. Bass**
Supreme Court of the United States. · 404 U.S. 336 · (1971)
... , or would welcome a cow that jumps over the moon, he would likely say "I ... would like to have a cat, a dog, or a cow that jumps over the moon." So it ... Cited By: 877

**Library Of Congress v. Shaw**
Supreme Court of the United States. · 478 U.S. 310 · (1986)

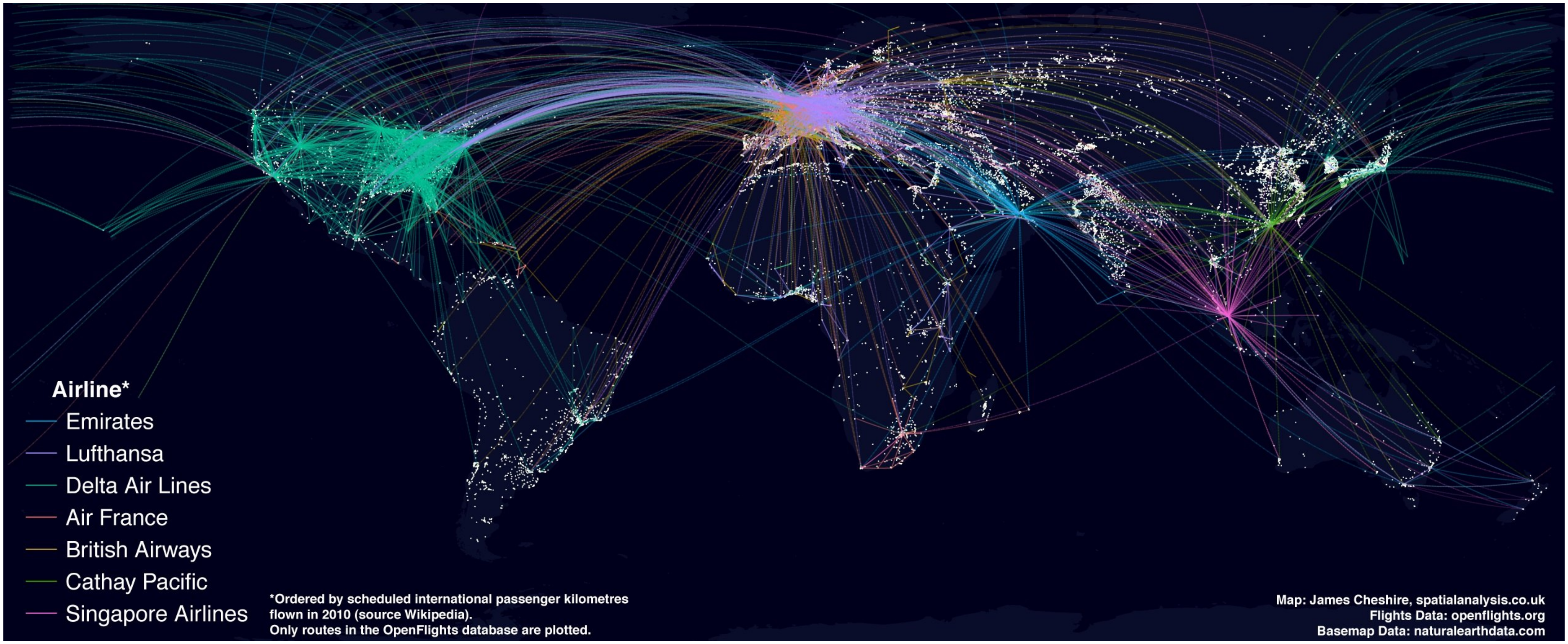
# Numerous Algorithms Exist

- Indegree Centrality
- PageRank
- HITS (Hyperlink-Induced Topic Search)
- Eigenvector Centrality
- And many, many, other stupefying algorithms

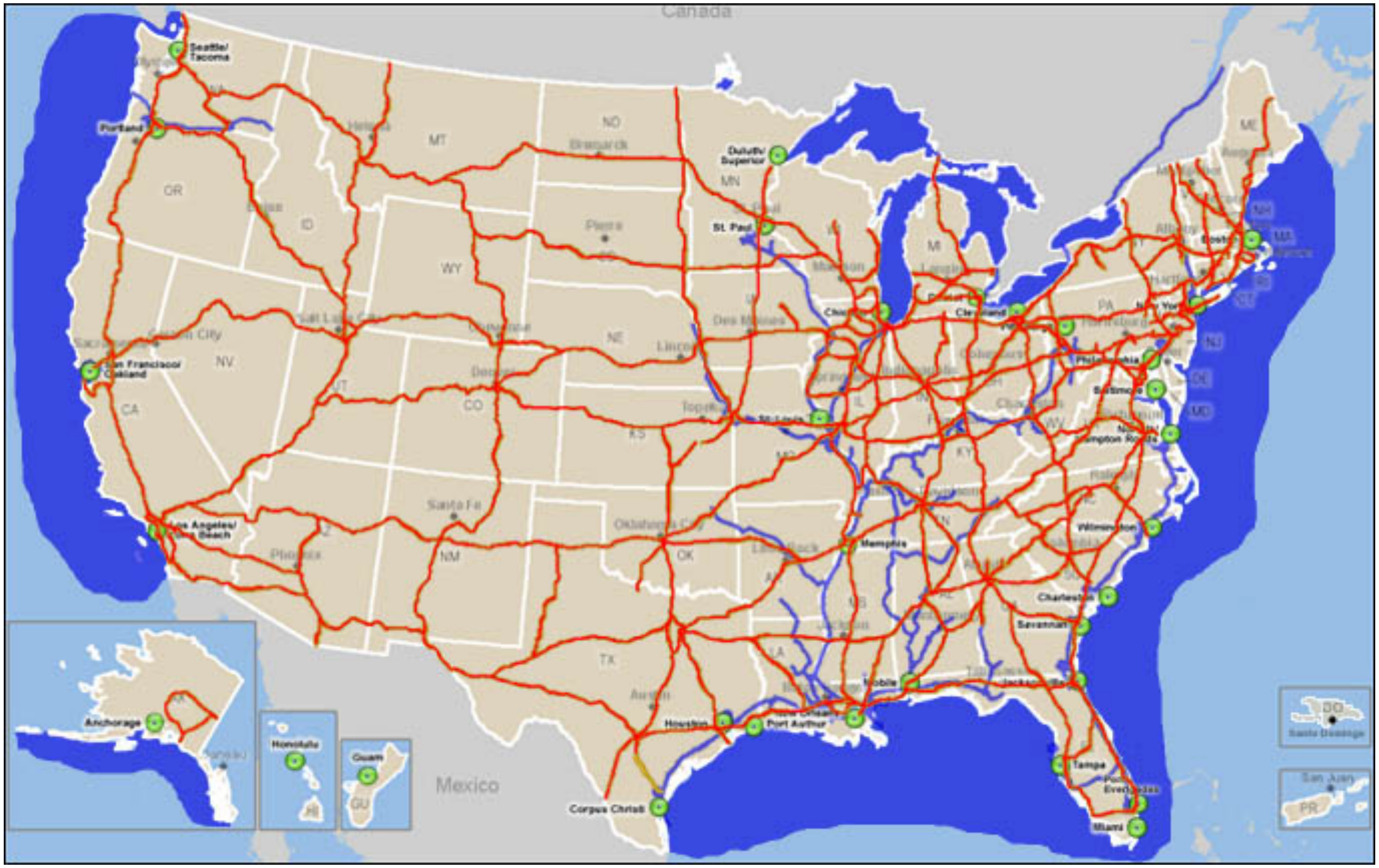
# The network of case citations is "scale-free"



# Scale-free network


Airline*
Emirates
Lufthansa
Delta Air Lines
Air France
British Airways
Cathay Pacific
Singapore Airlines
*Ordered by scheduled international passenger kilometres flown in 2010 (source Wikipedia).
Only routes in the OpenFlights database are plotted.
Map: James Cheshire, spatialanalysis.co.uk
Flights Data: openflights.org
Basemap Data: naturalearthdata.com

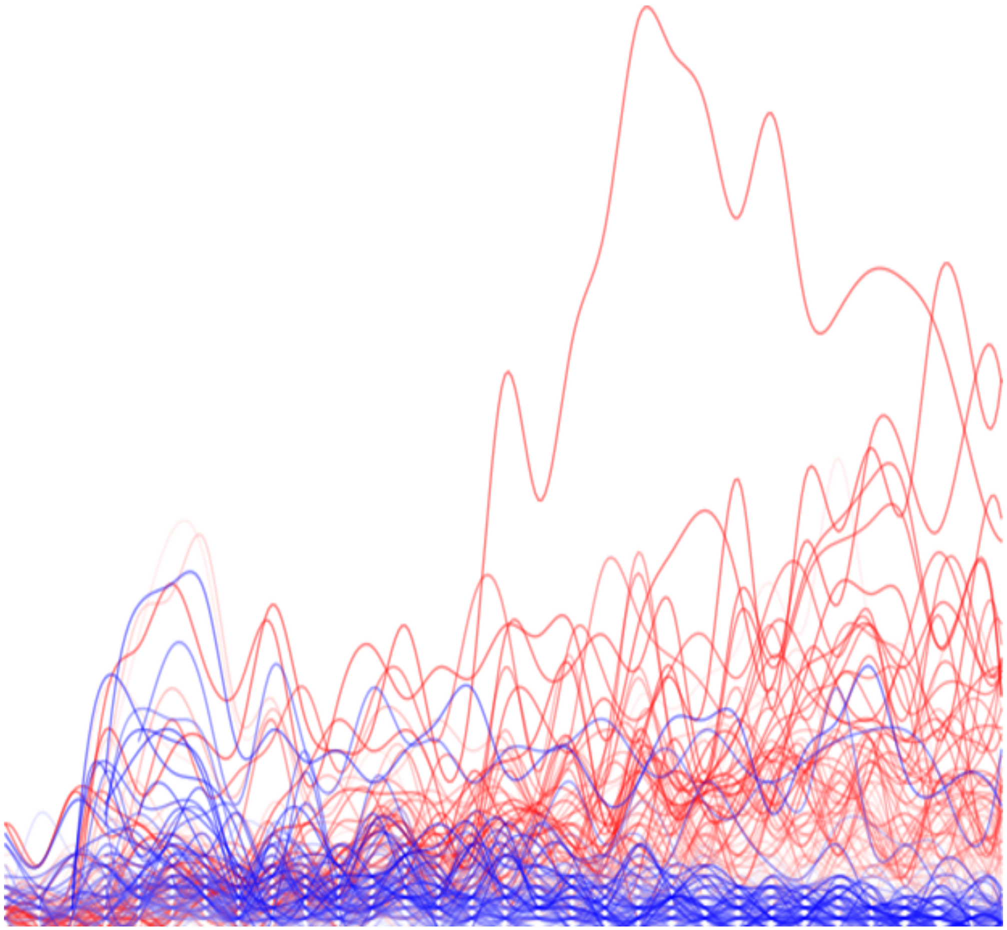
# Random network



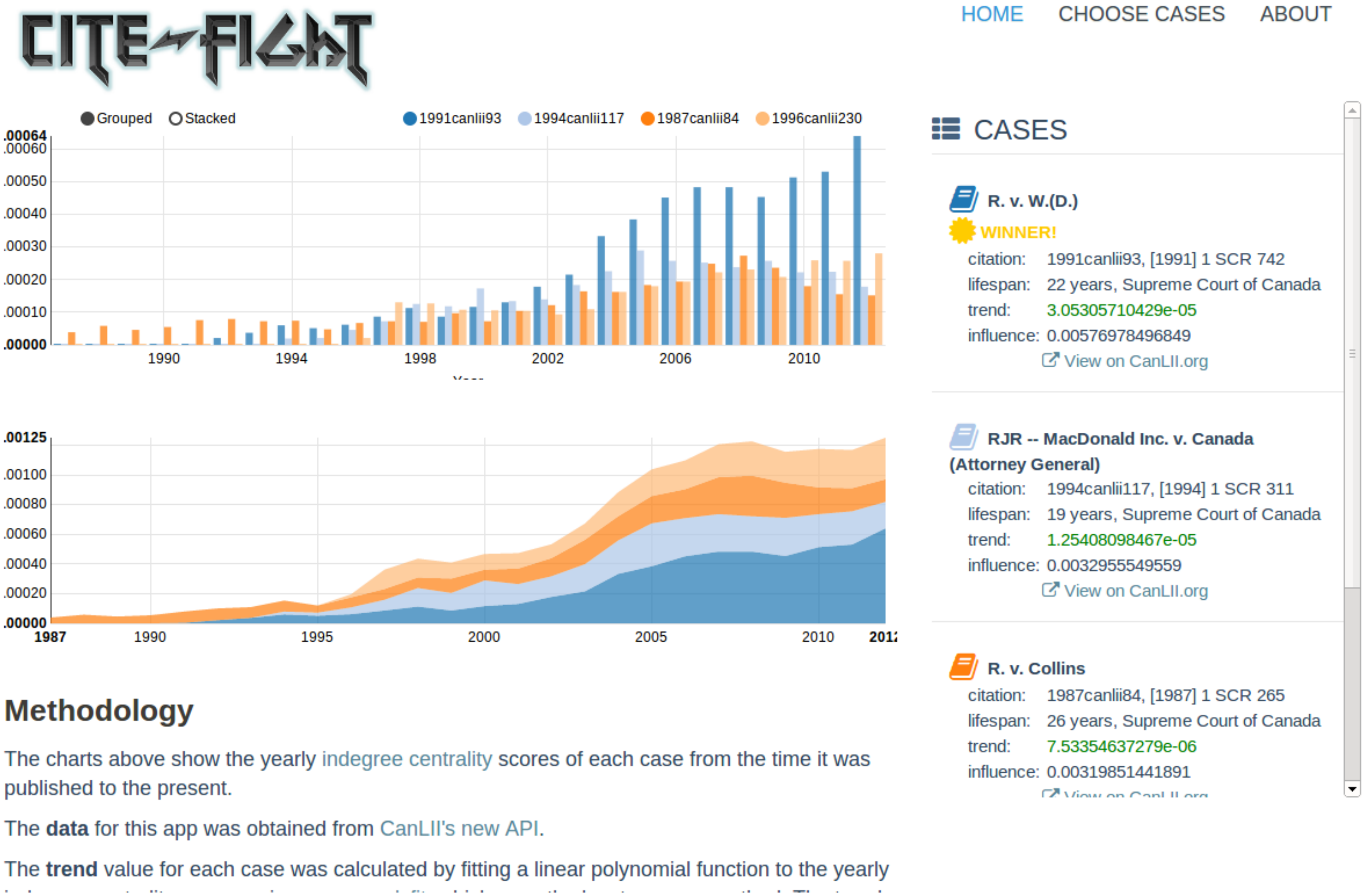
# Other important scale-free networks?

# 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

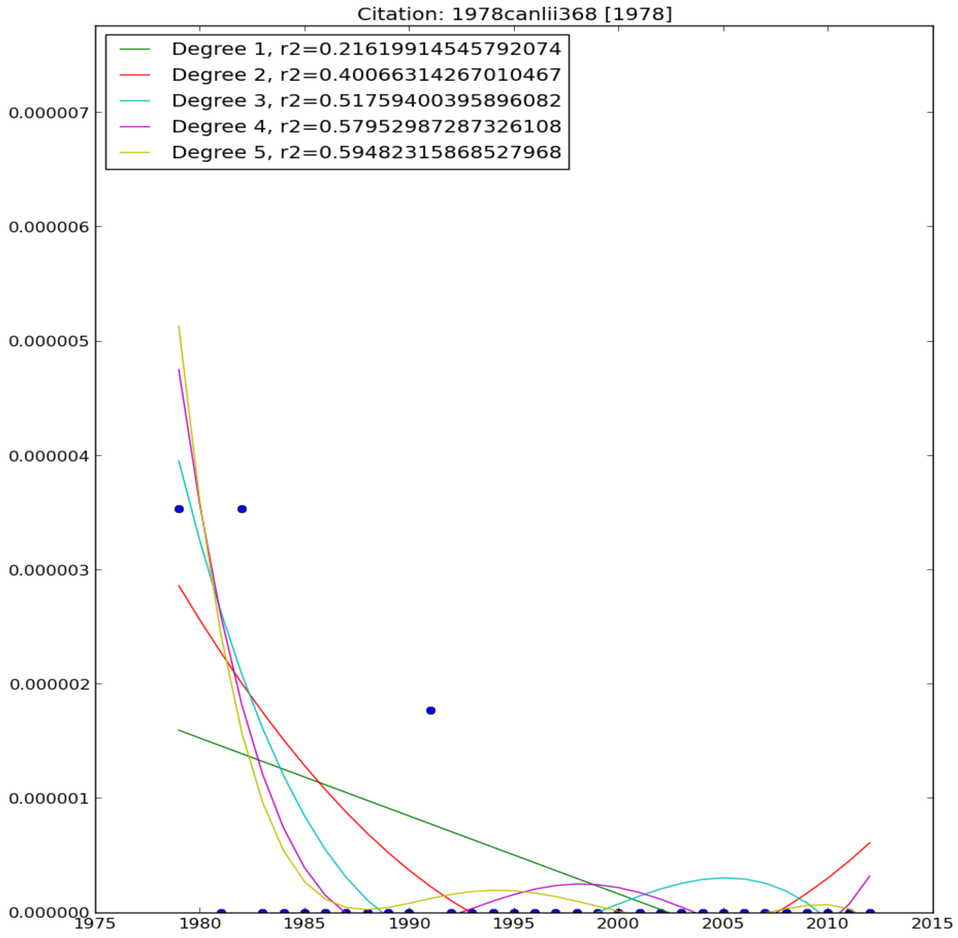
## CASES

**R. v. W.(D.)**
WINNER!
citation: 1991canlii93, [1991] 1 SCR 742
lifespan: 22 years, Supreme Court of Canada
trend: 3.05305710429e-05
influence: 0.00576978496849
View on CanLII.org

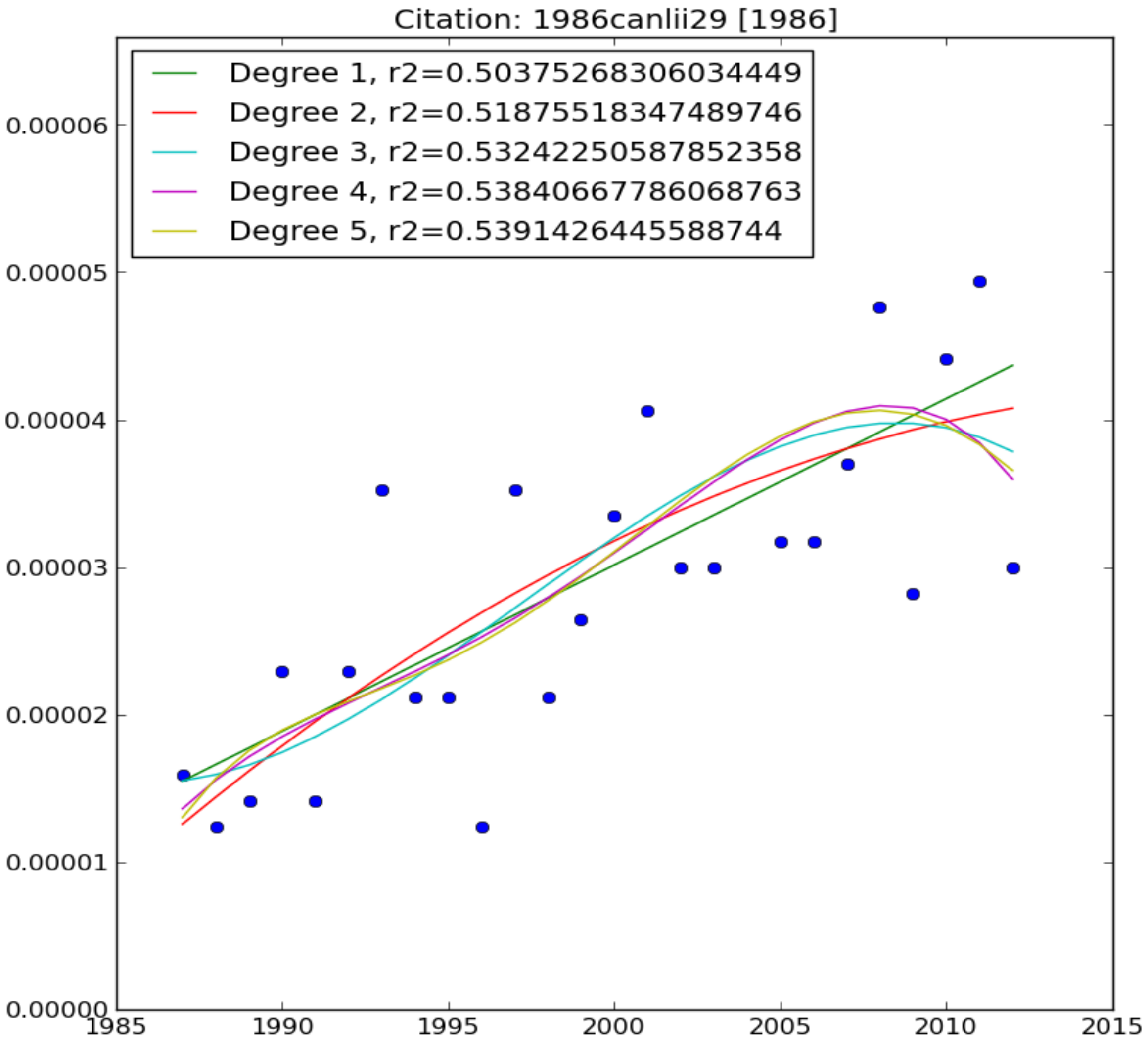
**RJR -- MacDonald Inc. v. Canada (Attorney General)**
citation: 1994canlii117, [1994] 1 SCR 311
lifespan: 19 years, Supreme Court of Canada
trend: 1.25408098467e-05
influence: 0.0032955549559
View on CanLII.org

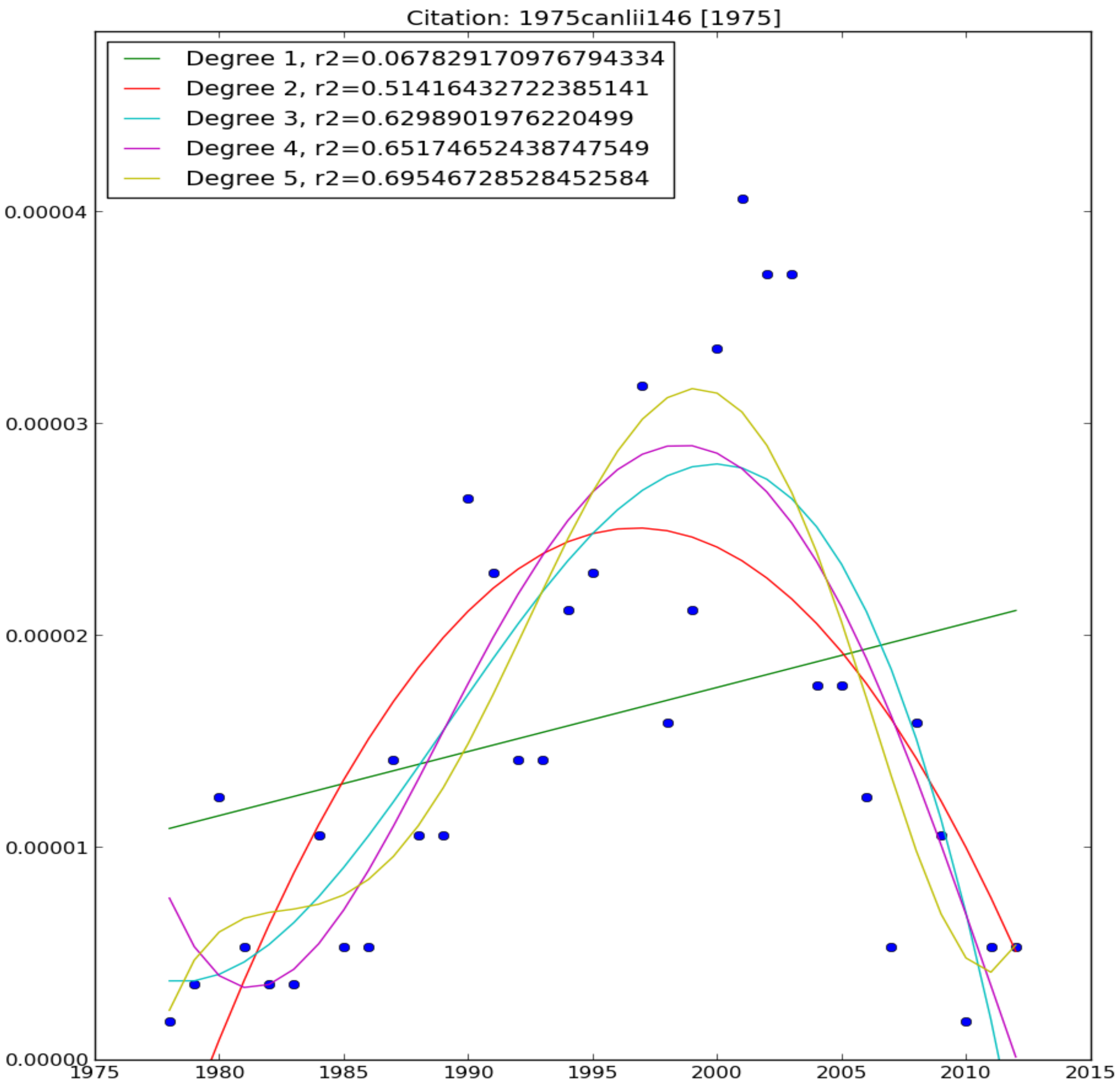
**R. v. Collins**
citation: 1987canlii84, [1987] 1 SCR 265
lifespan: 26 years, Supreme Court of Canada
trend: 7.53354637279e-06
influence: 0.00319851441891
View on CanLII.org

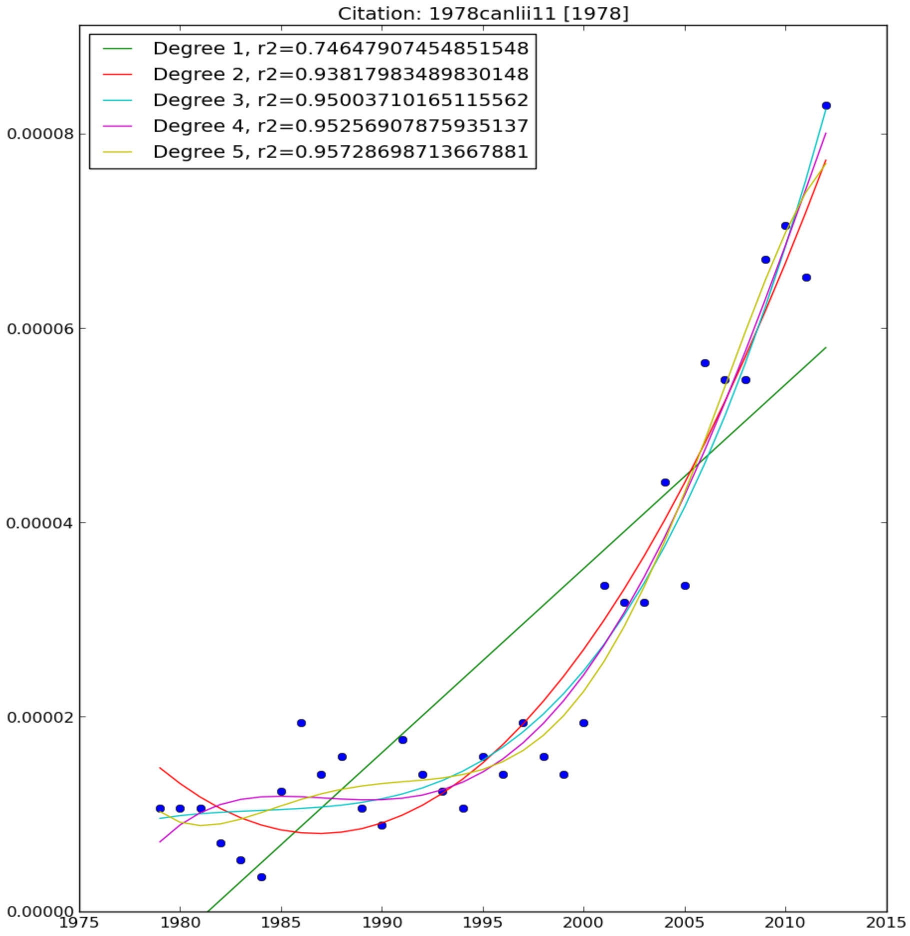
# You can fit curves to the data

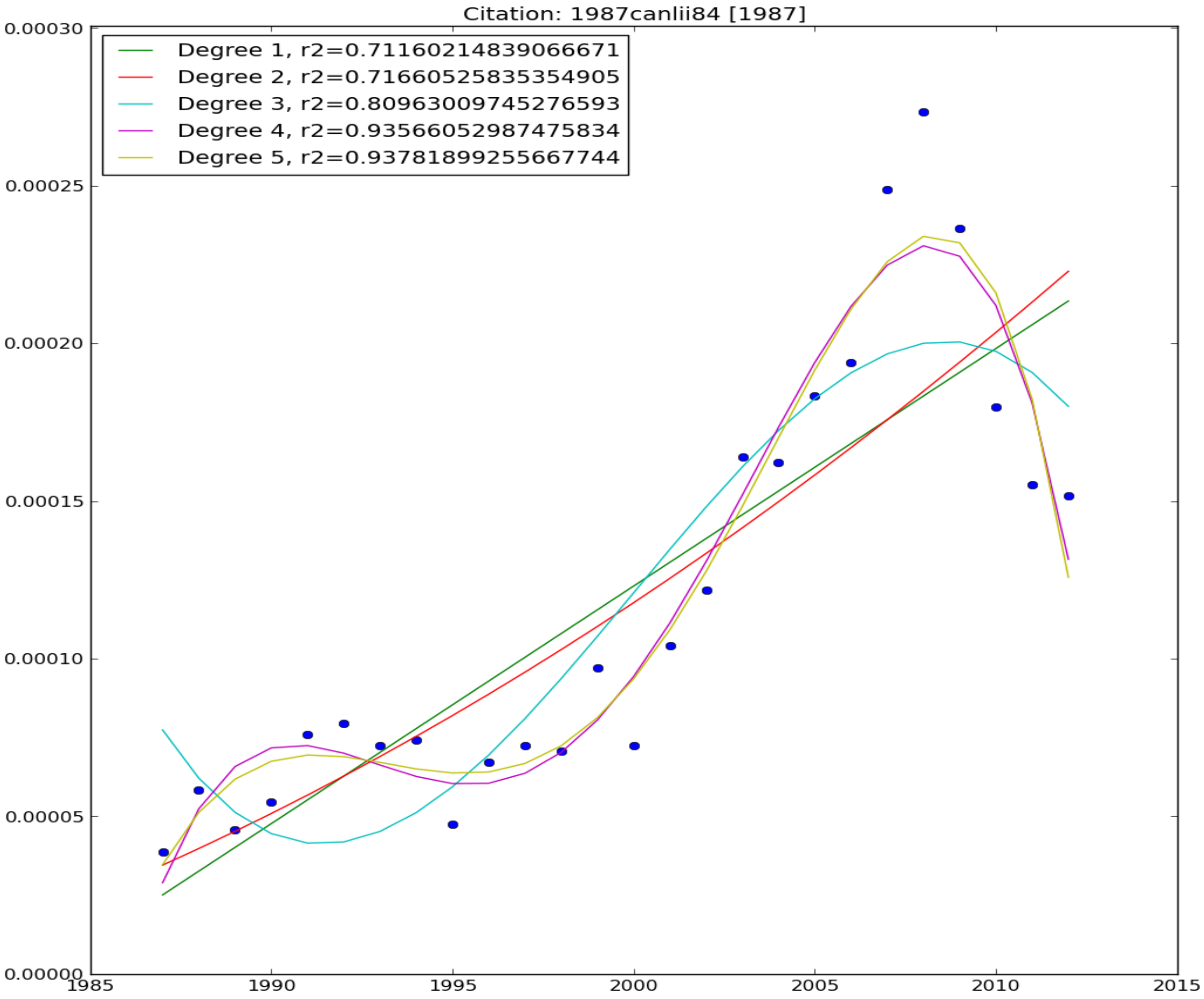


# The curves enable you to estimate things




Citation: 1975canlii146 [1975]
Degree 1, r2=0.067829170976794334
Degree 2, r2=0.51416432722385141
Degree 3, r2=0.6298901976220499
Degree 4, r2=0.65174652438747549
Degree 5, r2=0.69546728528452584
0.00004
0.00003
0.00002
0.00001
0.00000
1975
1980
1985
1990
1995
2000
2005
2010
2015

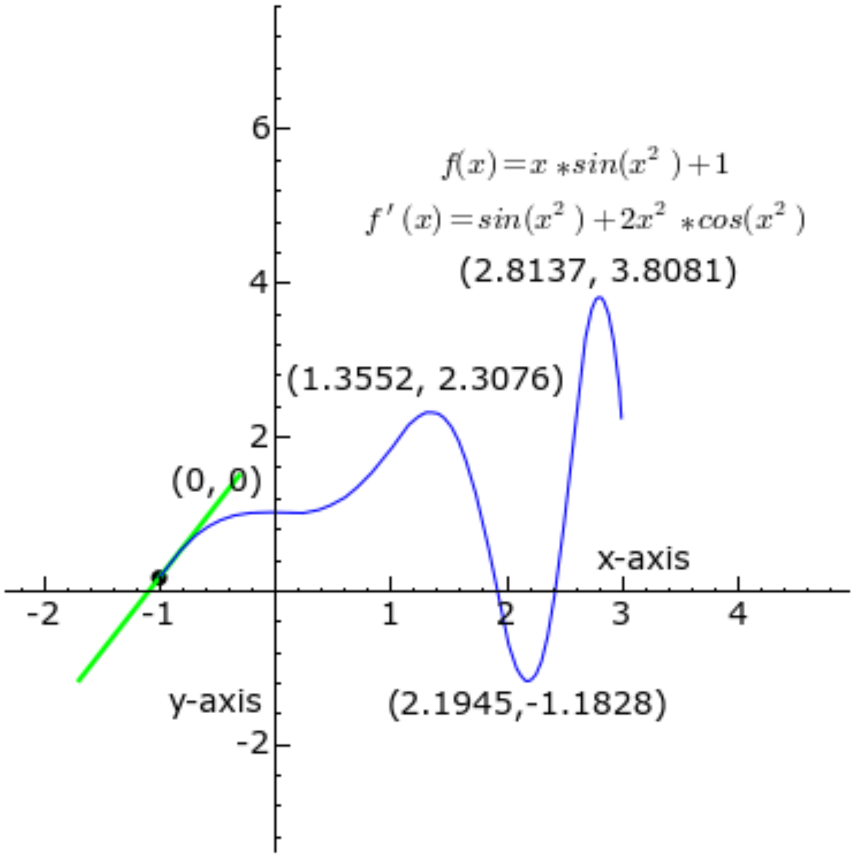

Citation: 1978canlii11 [1978]
Degree 1, r2=0.74647907454851548
Degree 2, r2=0.93817983489830148
Degree 3, r2=0.95003710165115562
Degree 4, r2=0.95256907875935137
Degree 5, r2=0.95728698713667881
0.00008
0.00006
0.00004
0.00002
0.00000
1975
1980
1985
1990
1995
2000
2005
2010
2015


Citation: 1987canlii84 [1987]
Degree 1, r2=0.71160214839066671
Degree 2, r2=0.71660525835354905
Degree 3, r2=0.80963009745276593
Degree 4, r2=0.93566052987475834
Degree 5, r2=0.93781899255667744
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
1985
1990
1995
2000
2005
2010
2015

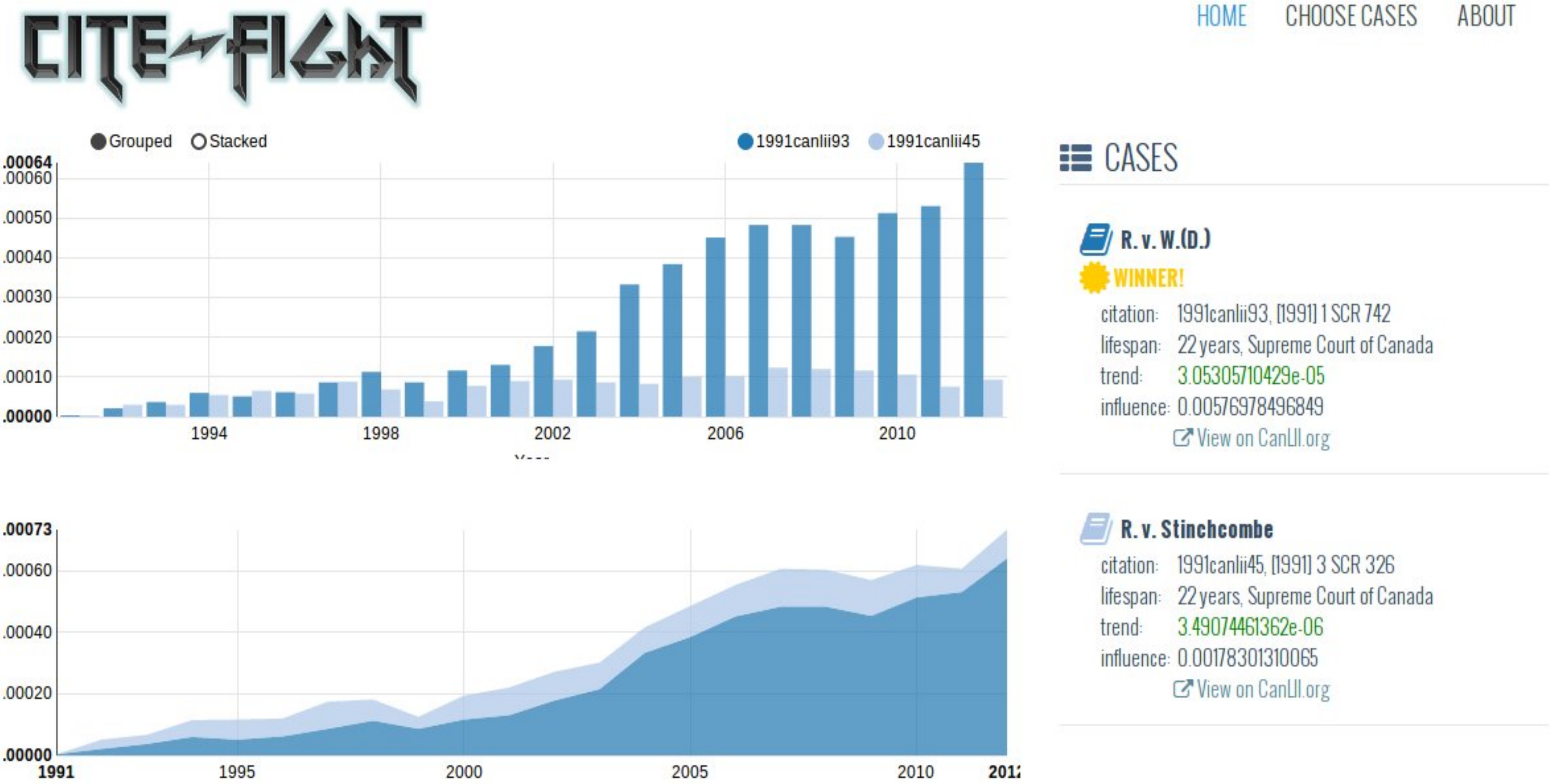
# What things?

- 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?



# Which of these cases has had a greater cumulative influence over time?



CASES

R. v. W.(D.)

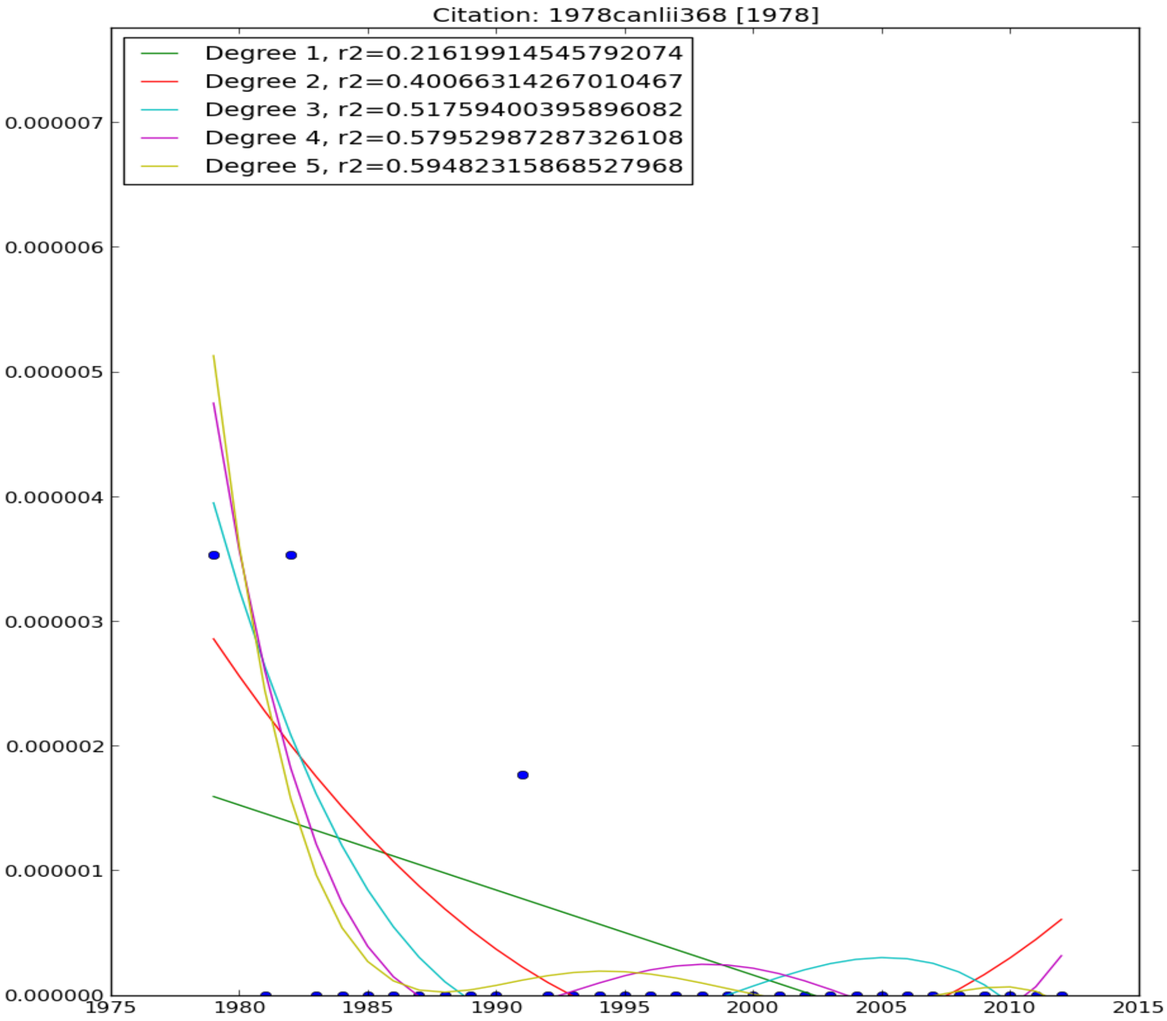
WINNER!

citation: 1991canlii93, [1991] 1 SCR 742
lifespan: 22 years, Supreme Court of Canada
trend: 3.05305710429e-05
influence: 0.00576978496849
View on CanLII.org

R. v. Stinchcombe

citation: 1991canlii45, [1991] 3 SCR 326
lifespan: 22 years, Supreme Court of Canada
trend: 3.49074461362e-06
influence: 0.00178301310065
View on CanLII.org

# 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

# This

“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.).”

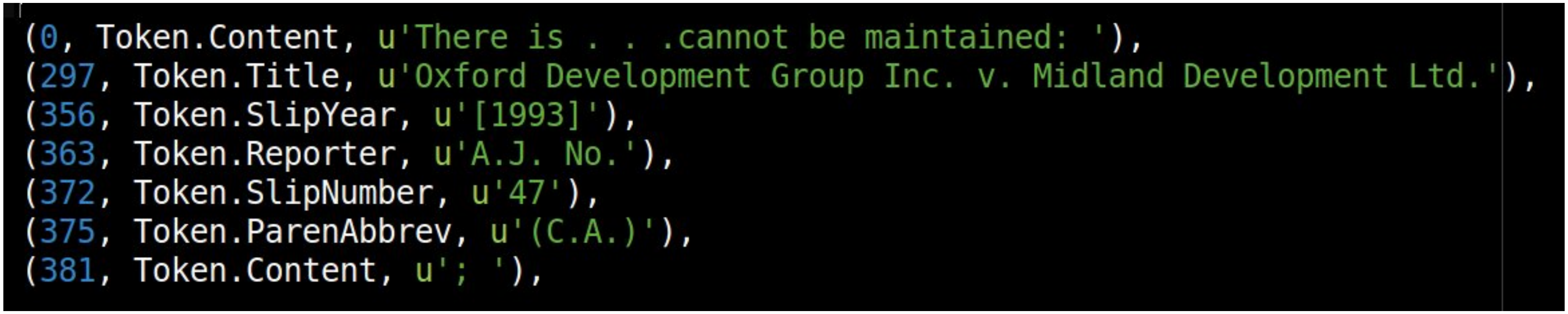
# This

"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.)."

# This

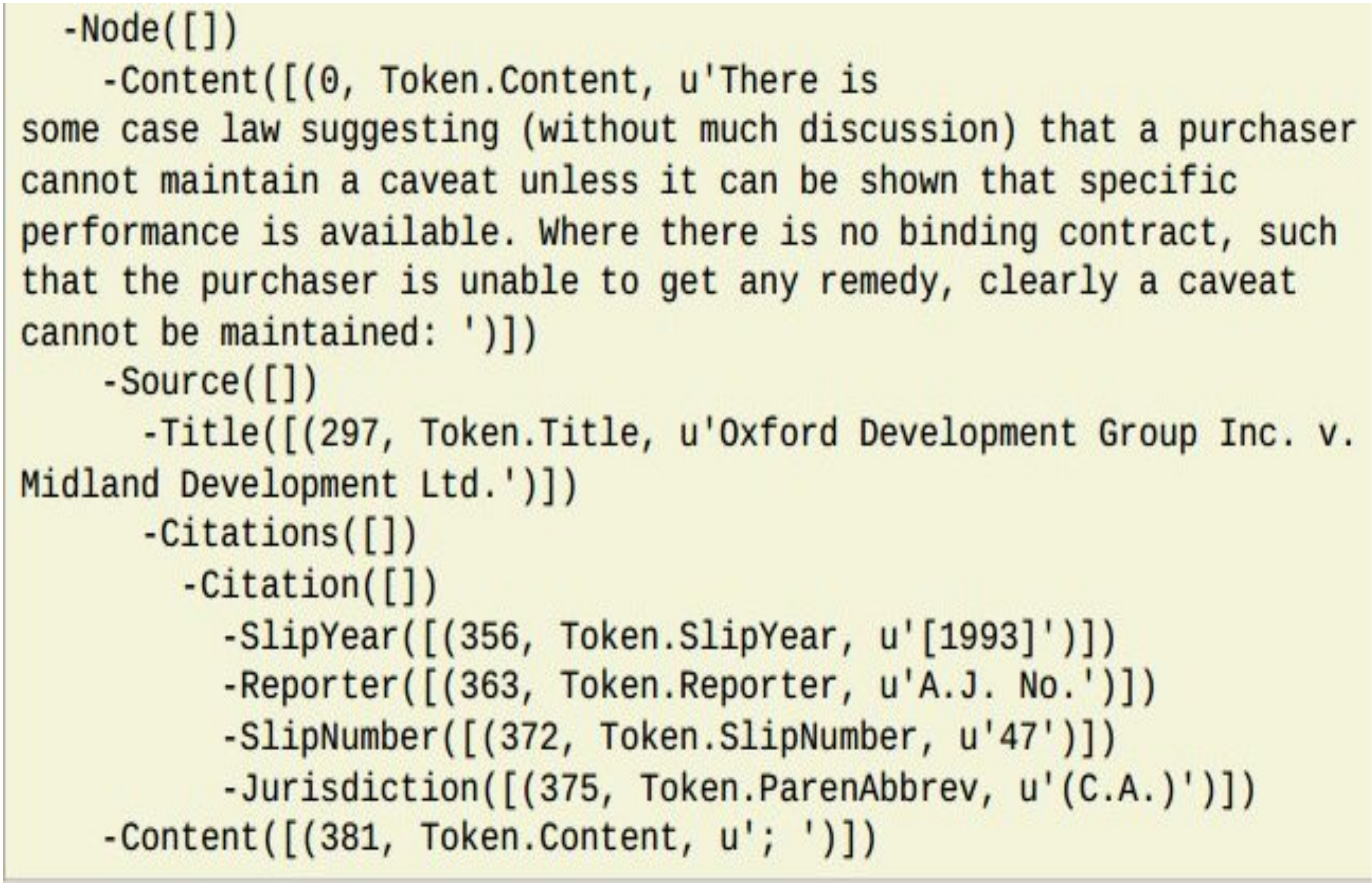
"There is some case law suggesting (without much discussion) that a purchaser cannot maintain a caveat unless it can be shown that specific performance is available. Where there is no binding contract, such that the purchaser is unable to get any remedy, clearly a caveat cannot be maintained: Oxford Development Group Inc. v. Midland Development Ltd., [1993] A.J. No. 47 (C.A.)."

# Becomes this



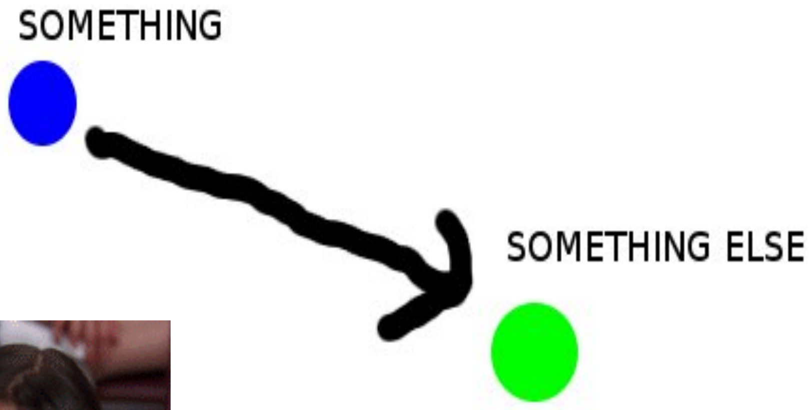
```
(0, Token.Content, u'There is . . .cannot be maintained: '),
(297, Token.Title, u'Oxford Development Group Inc. v. Midland Development Ltd.'),
(356, Token.SlipYear, u'[1993]'),
(363, Token.Reporter, u'A.J. No.'),
(372, Token.SlipNumber, u'47'),
(375, Token.ParenAbbrev, u'(C.A.)'),
(381, Token.Content, u'; '),
```

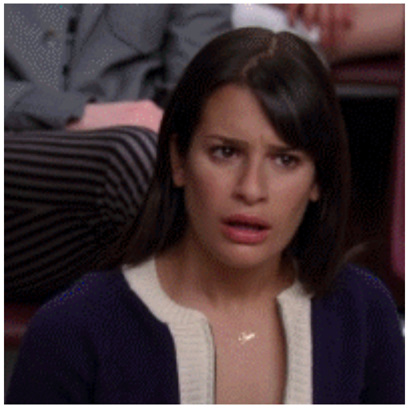
# Then This



```
  -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!





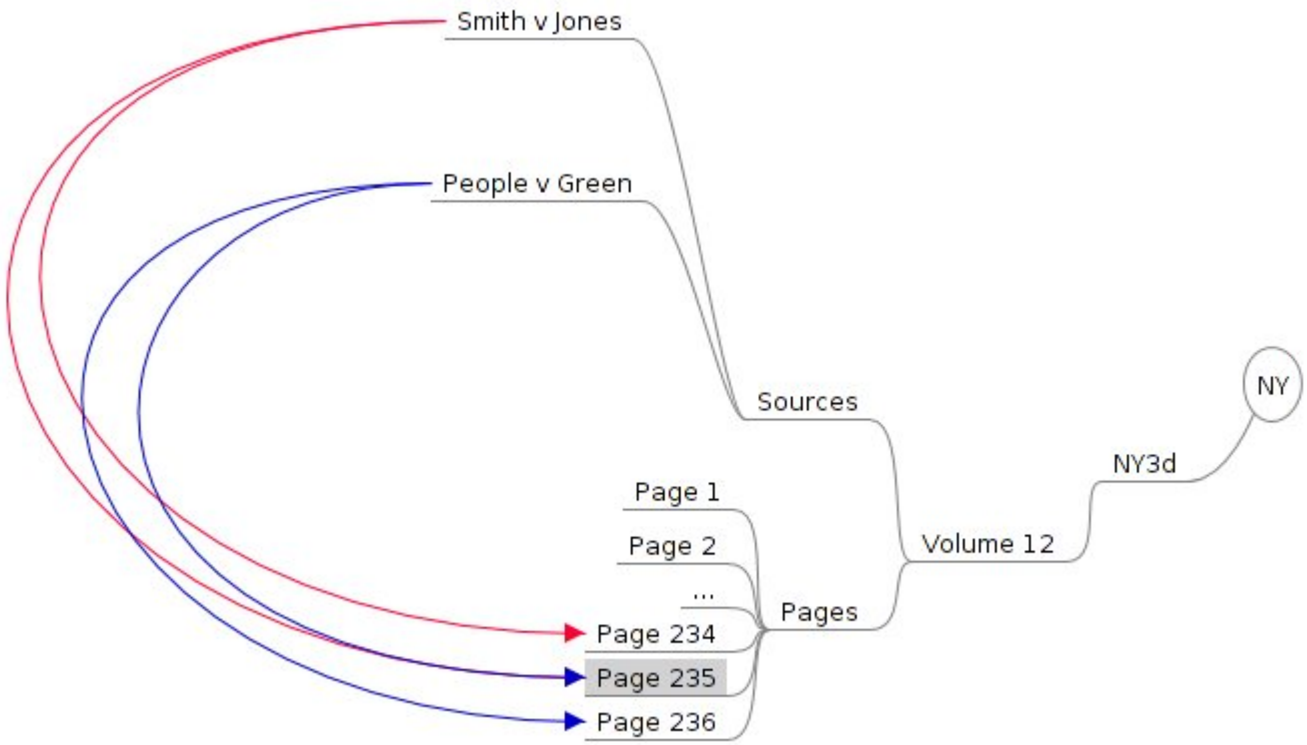
# Citation Data Should Probably Be Stored as a Graph

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

# Really Hard

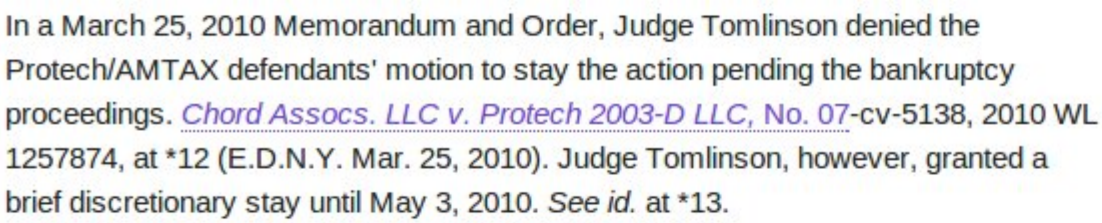
- Resolving citations back to sources
- Why:
  - Volume pages are not unique identifiers
  - Titles aren't unique identifiers either
  - An alarming percentage of citations contain typos

# Volume Pages aren't unique IDs

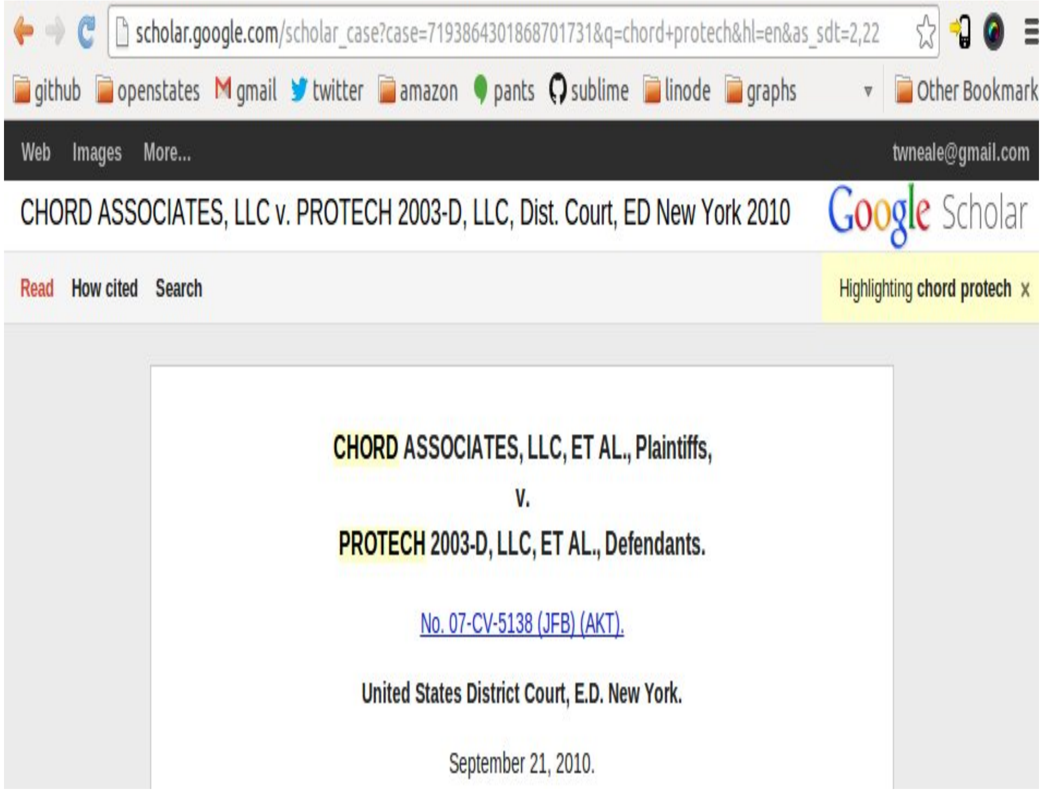


# Comparing Titles

- Is unreliable
- Even hard for



In a March 25, 2010 Memorandum and Order, Judge Tomlinson denied the Protech/AMTAX defendants' motion to stay the action pending the bankruptcy proceedings. *Chord Assocs. LLC v. Protech 2003-D LLC*, No. 07-cv-5138, 2010 WL 1257874, at *12 (E.D.N.Y. Mar. 25, 2010). Judge Tomlinson, however, granted a brief discretionary stay until May 3, 2010. *See id.* at *13.



CHORD ASSOCIATES, LLC v. PROTECH 2003-D, LLC, Dist. Court, ED New York 2010

CHORD ASSOCIATES, LLC, ET AL., Plaintiffs,

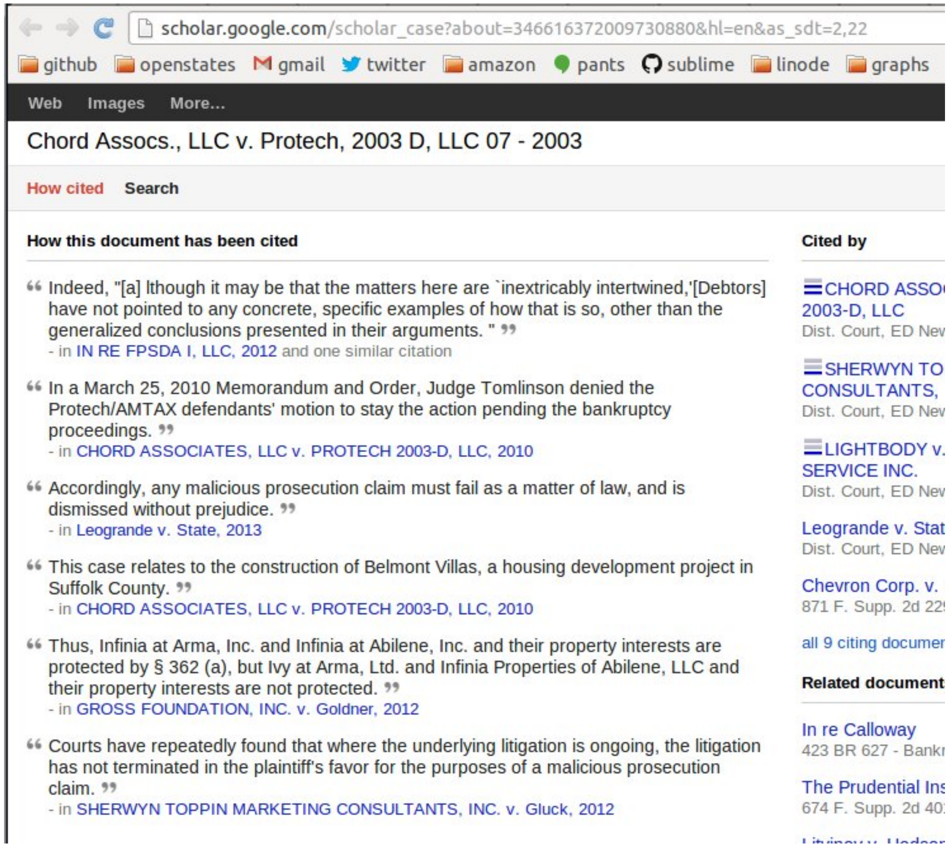
v.

PROTECH 2003-D, LLC, ET AL., Defendants.

No. 07-CV-5138 (JFB) (AKT).

United States District Court, E.D. New York.

September 21, 2010.



Chord Assocs., LLC v. Protech, 2003 D, LLC 07 - 2003

How this document has been cited

" Indeed, "[a] lthough it may be that the matters here are `inextricably intertwined,'[Debtors] have not pointed to any concrete, specific examples of how that is so, other than the generalized conclusions presented in their arguments. " "
- in IN RE FPSDA I, LLC, 2012 and one similar citation

" In a March 25, 2010 Memorandum and Order, Judge Tomlinson denied the Protech/AMTAX defendants' motion to stay the action pending the bankruptcy proceedings. "
- in CHORD ASSOCIATES, LLC v. PROTECH 2003-D, LLC, 2010

" Accordingly, any malicious prosecution claim must fail as a matter of law, and is dismissed without prejudice. "
- in Leogrande v. State, 2013

" This case relates to the construction of Belmont Villas, a housing development project in Suffolk County. "
- in CHORD ASSOCIATES, LLC v. PROTECH 2003-D, LLC, 2010

" Thus, Infinia at Arma, Inc. and Infinia at Abilene, Inc. and their property interests are protected by § 362 (a), but Ivy at Arma, Ltd. and Infinia Properties of Abilene, LLC and their property interests are not protected. "
- in GROSS FOUNDATION, INC. v. Goldner, 2012

" Courts have repeatedly found that where the underlying litigation is ongoing, the litigation has not terminated in the plaintiff's favor for the purposes of a malicious prosecution claim. "
- in SHERWYN TOPPIN MARKETING CONSULTANTS, INC. v. Gluck, 2012

Cited by

CHORD ASSO
2003-D, LLC
Dist. Court, ED New

SHERWYN TO
CONSULTANTS,
Dist. Court, ED New

LIGHTBODY v.
SERVICE INC.
Dist. Court, ED New

Leogrande v. Stat
Dist. Court, ED New

Chevron Corp. v.
871 F. Supp. 2d 22

all 9 citing documer

Related document

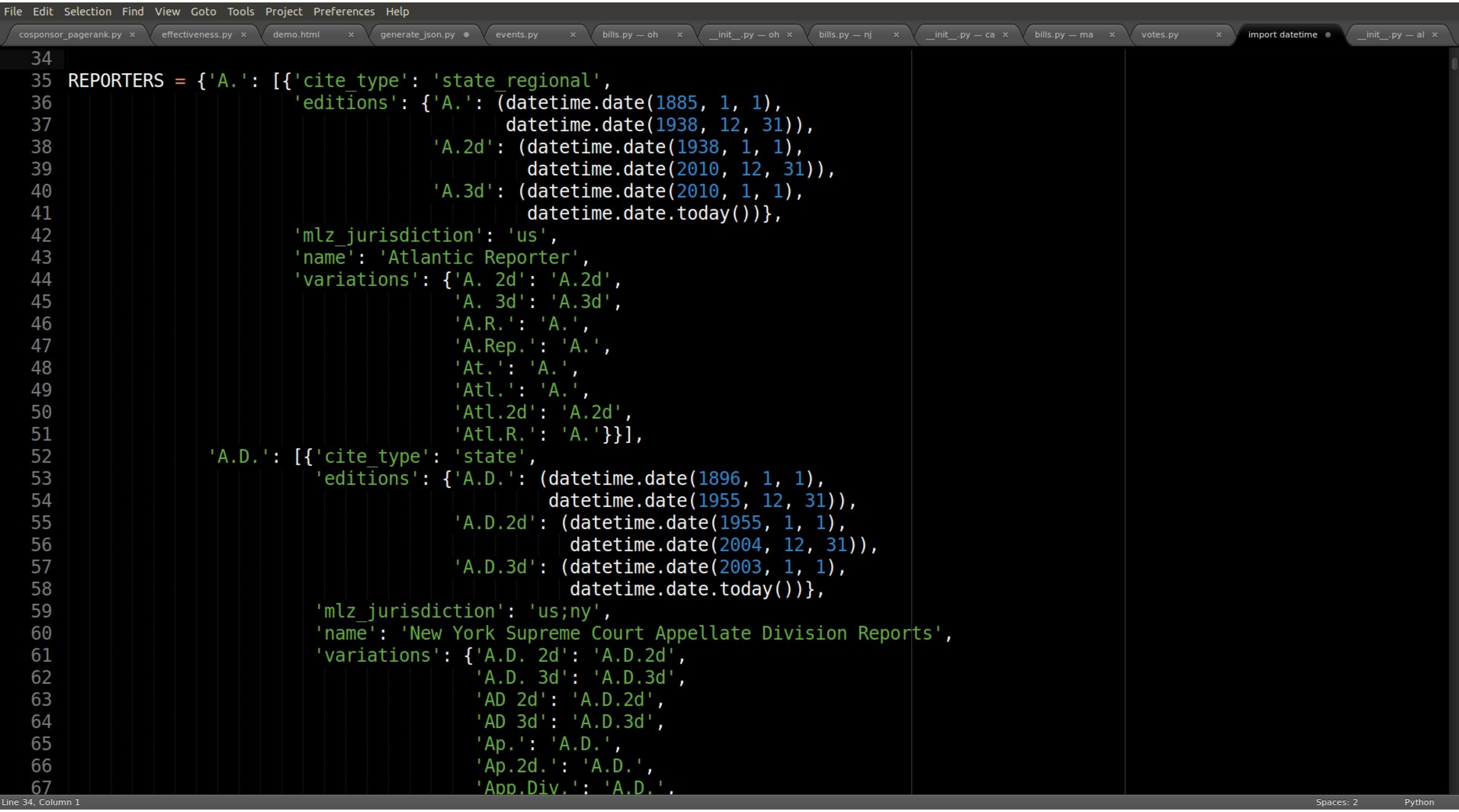
In re Calloway
423 BR 627 - Bankr

The Prudential Ins
674 F. Supp. 2d 40

# 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?

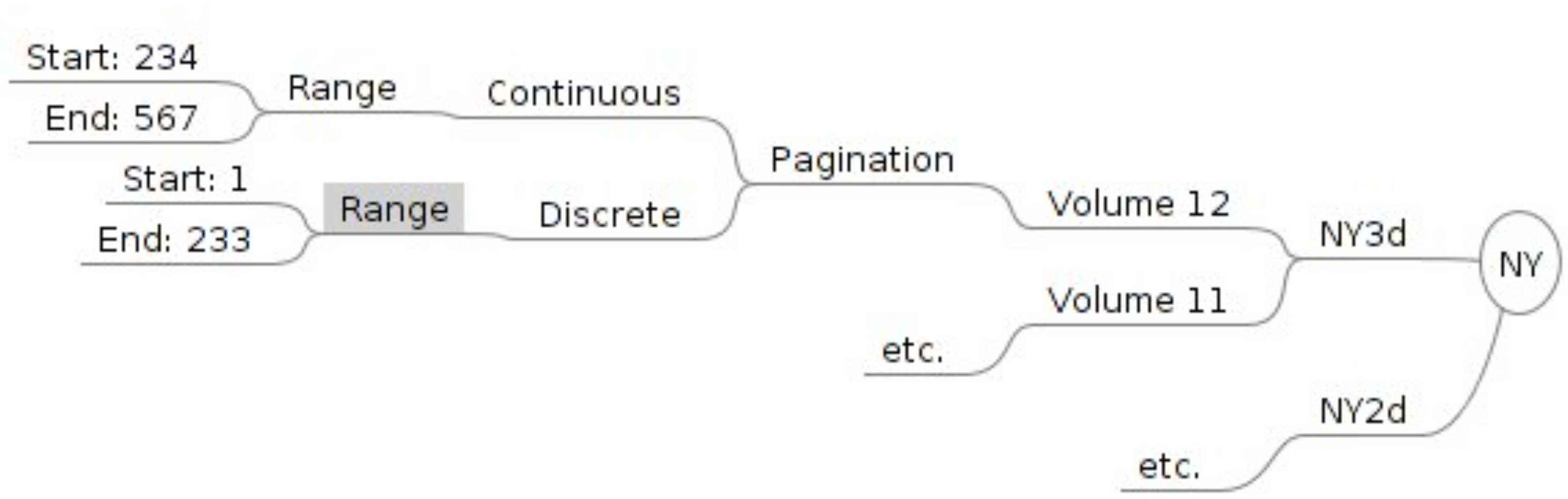
# Need Publication Metadata



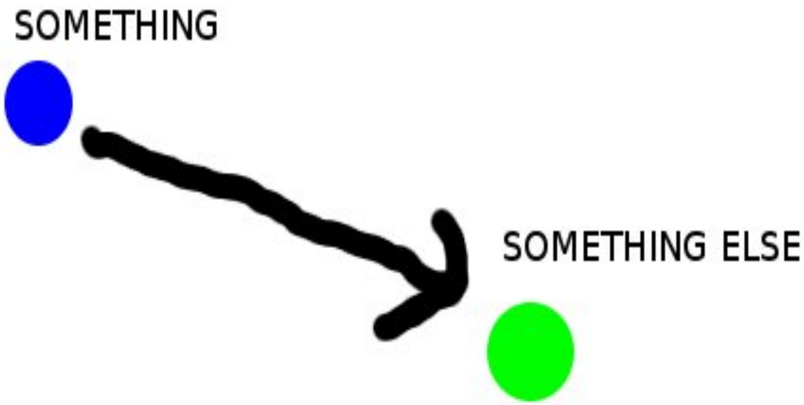
```

REPORTERS = {'A.': [{'cite_type': 'state_regional',
                     'editions': {'A.': (datetime.date(1885, 1, 1),
                                         datetime.date(1938, 12, 31)),
                                  'A.2d': (datetime.date(1938, 1, 1),
                                           datetime.date(2010, 12, 31)),
                                  'A.3d': (datetime.date(2010, 1, 1),
                                           datetime.date.today())},
                     'mlz_jurisdiction': 'us',
                     'name': 'Atlantic Reporter',
                     'variations': {'A. 2d': 'A.2d',
                                    'A. 3d': 'A.3d',
                                    'A.R.': 'A.',
                                    'A.Rep.': 'A.',
                                    'At.': 'A.',
                                    'Atl.': 'A.',
                                    'Atl.2d': 'A.2d',
                                    'Atl.R.': 'A.'}}],
             'A.D.': [{'cite_type': 'state',
                       'editions': {'A.D.': (datetime.date(1896, 1, 1),
                                             datetime.date(1955, 12, 31)),
                                    'A.D.2d': (datetime.date(1955, 1, 1),
                                               datetime.date(2004, 12, 31)),
                                    'A.D.3d': (datetime.date(2003, 1, 1),
                                               datetime.date.today())},
                       'mlz_jurisdiction': 'us;ny',
                       'name': 'New York Supreme Court Appellate Division Reports',
                       'variations': {'A.D. 2d': 'A.D.2d',
                                      'A.D. 3d': 'A.D.3d',
                                      'AD 2d': 'A.D.2d',
                                      'AD 3d': 'A.D.3d',
                                      'Ap.': 'A.D.',
                                      'Ap.2d.': 'A.D.',
                                      'App.Div.': 'A.D.',
```

# It might look like this



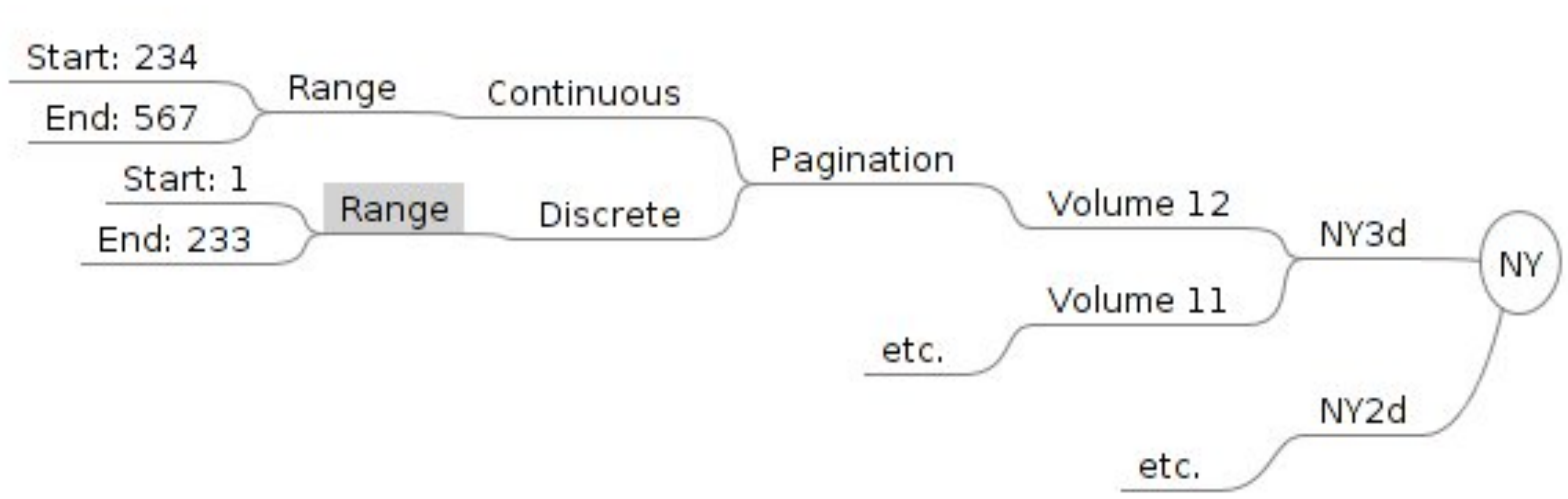
# Wait a minute...



SOMETHING

SOMETHING ELSE

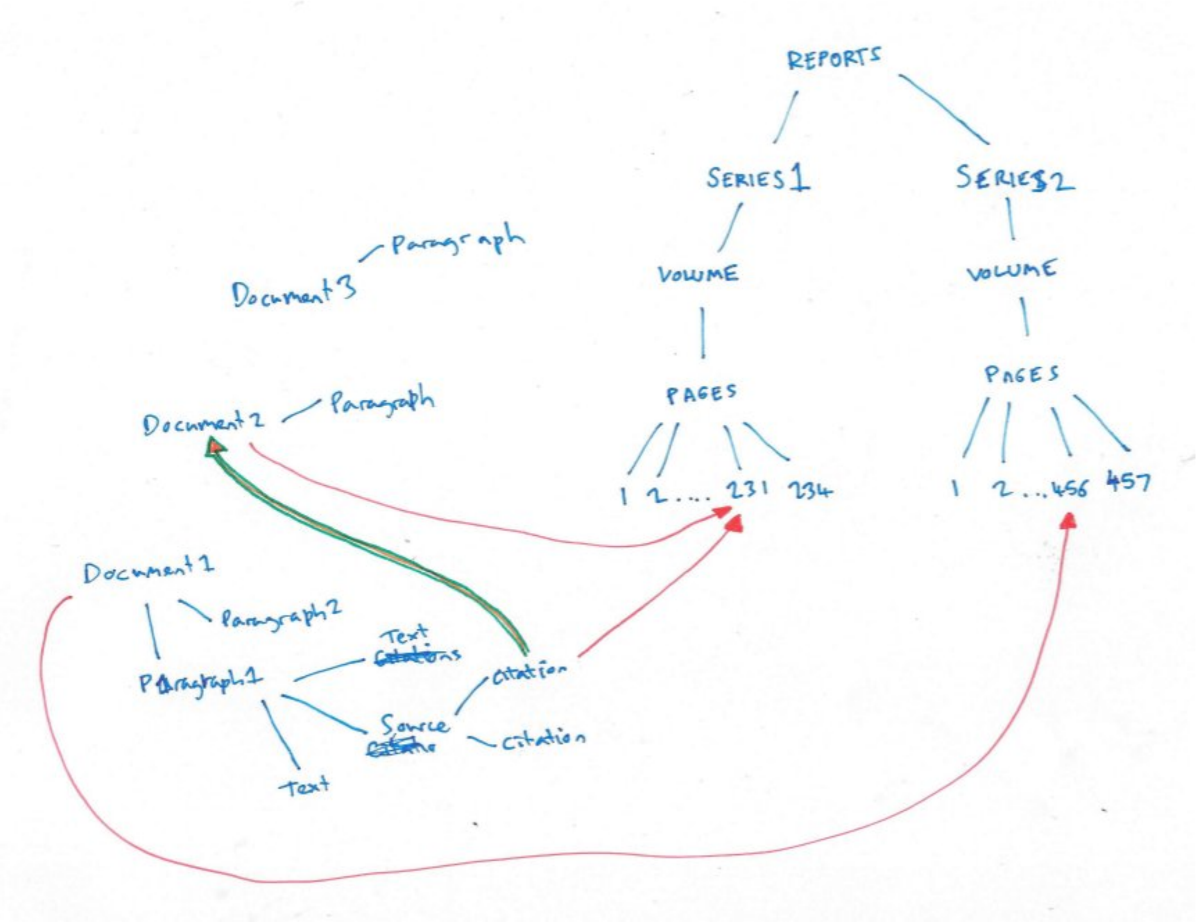
# It might look like this



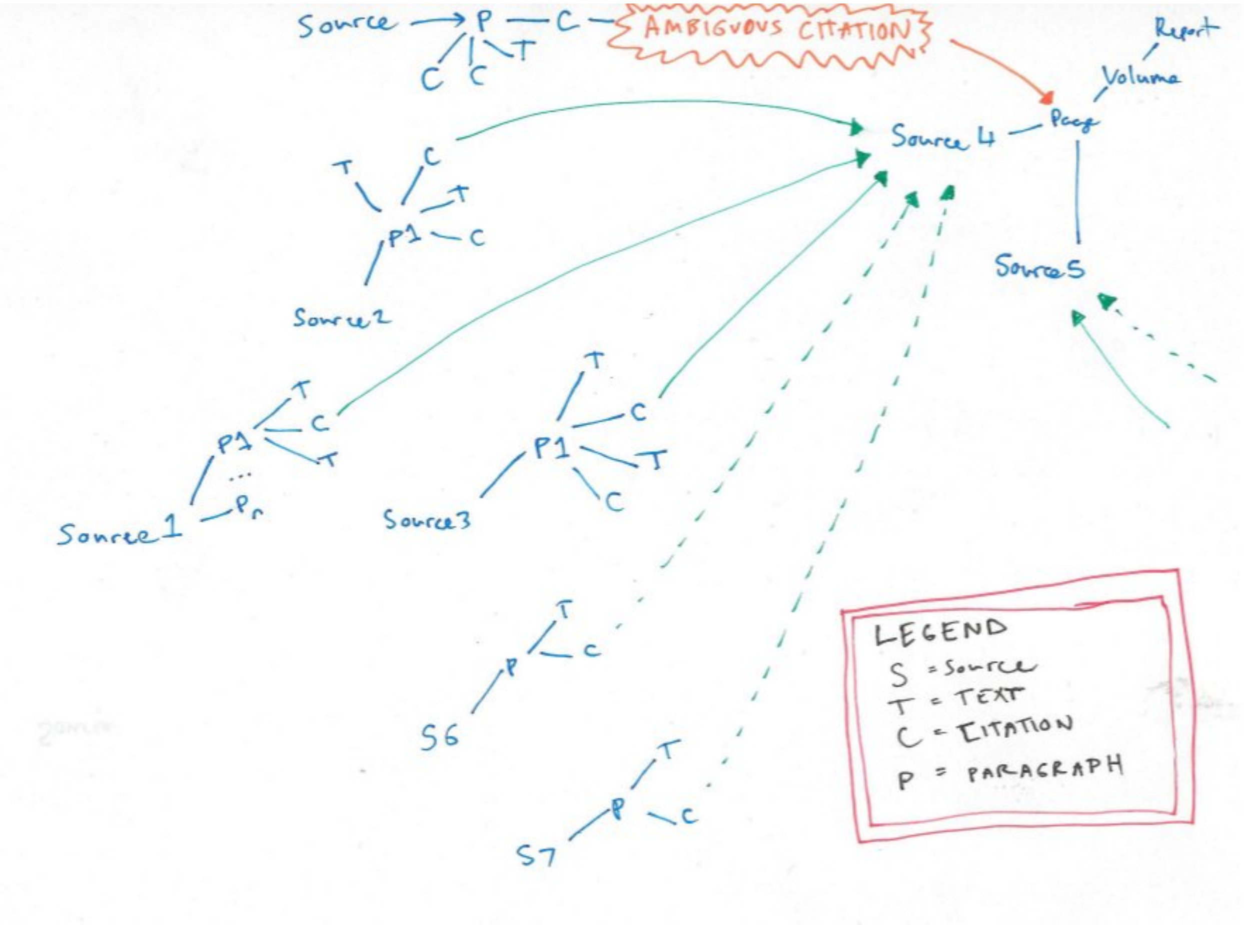
# No Thom

None of this is useful. Take a yellow pad back to your office and create some value for once.

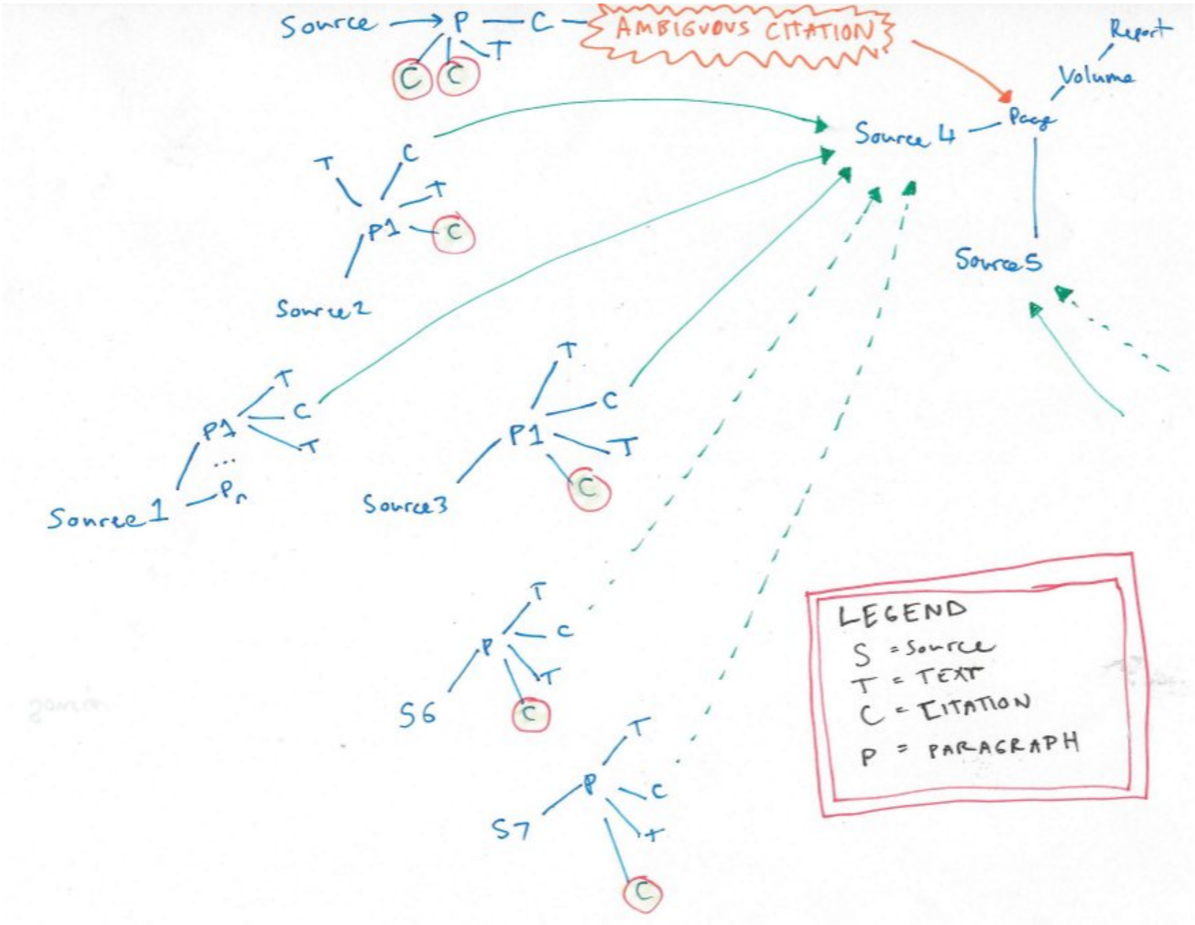
# An Overcomplicated Hypo



# 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

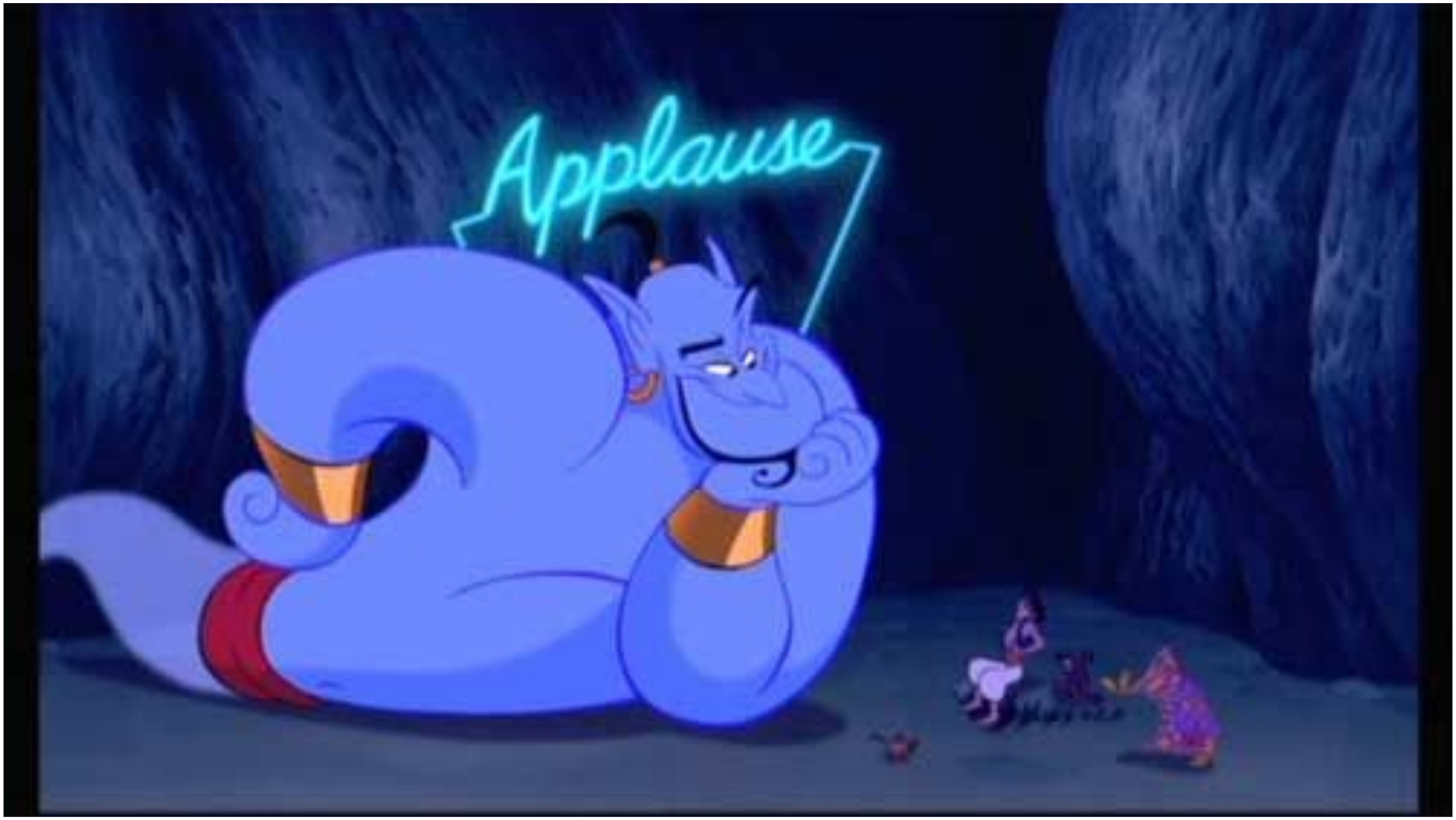
# The Technology

- 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



# Citation Network Analysis

Thom Neale
Twitter: @twneale
Github: twneale

twneale.github.io/citation-network-analysis