Patent Main Path Analysis with Arc Weights Adjusted by Classification Similarity

Max Kuan
National Taiwan University of Science and Technology

Center of Science and Technology Informetrics

Outlines

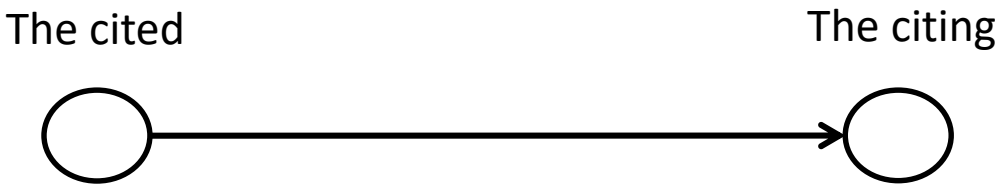
- **Background**
- Methodology
- Empirical Result
- Conclusion

Main path analysis

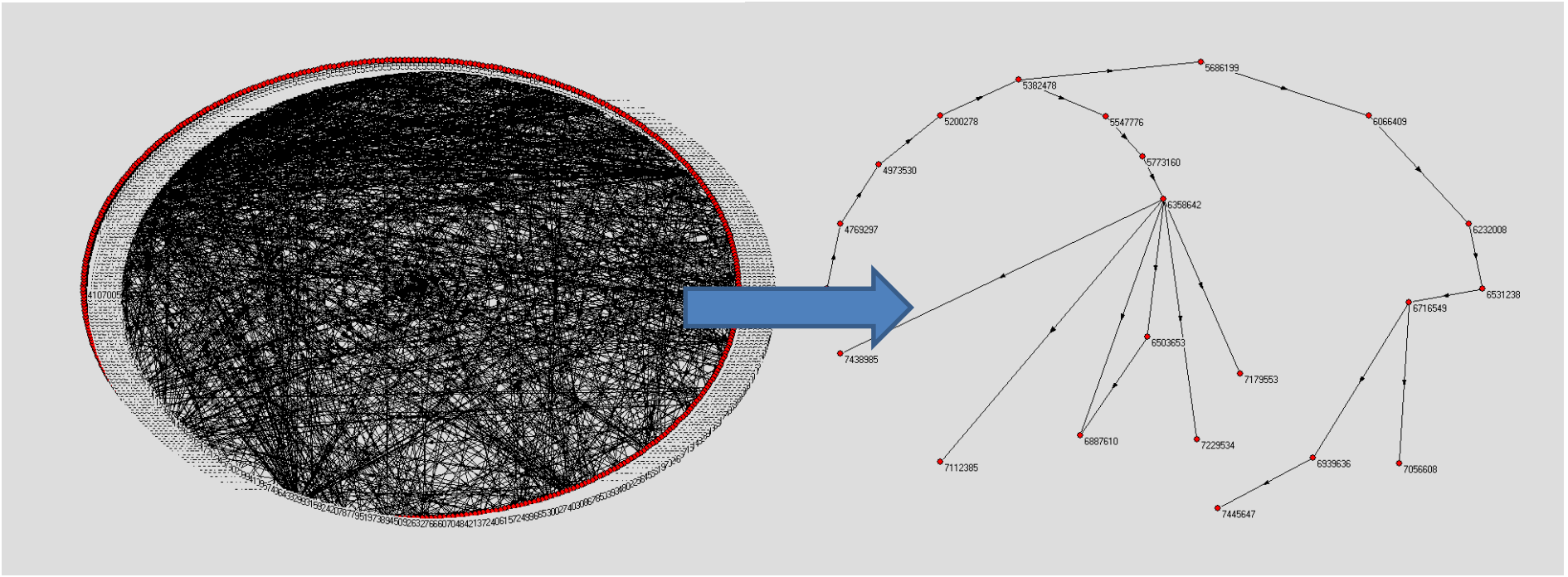
- A methodology from social network analysis
- Originally aimed to determine the major development trajectory in a scientific field
- Has been applied to patents in detecting technological changes, and finding trajectories of technological development

Basic Assumption

Citation≈Knowledge flow



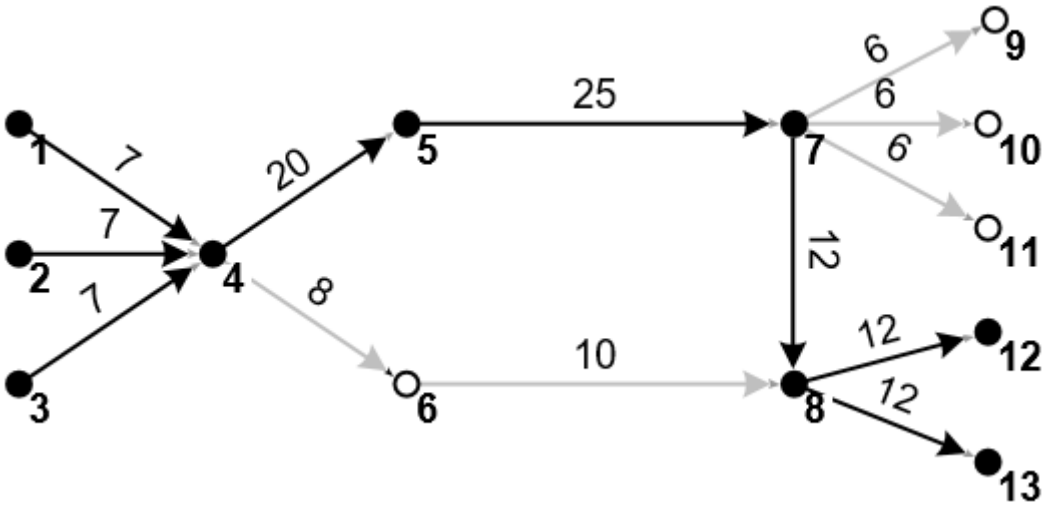
An Example



Analysis procedure

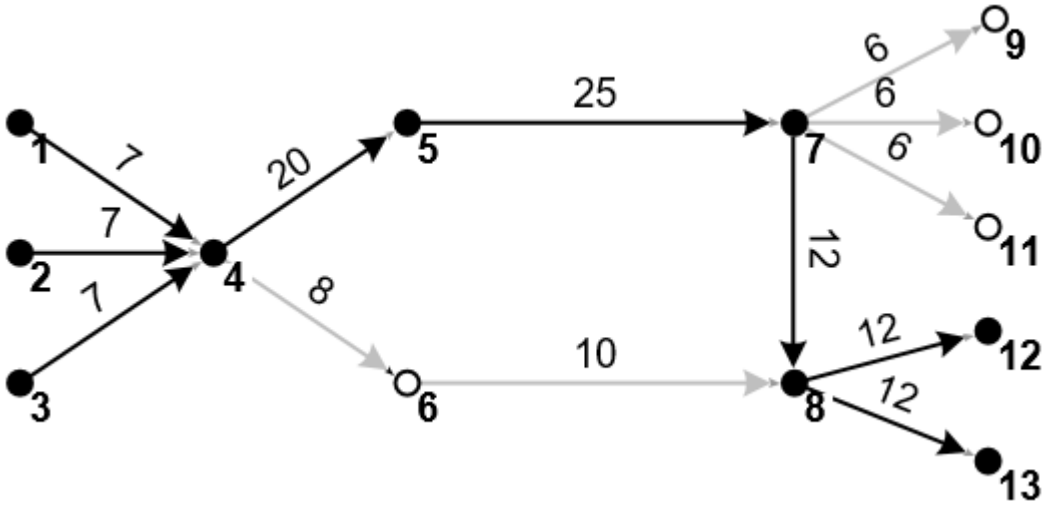
1. A citation network is constructed
2. A weight for each arc is assigned based on its traversal count
   - There are various algorithms
3. A series of connected arcs across the network is determined as a representative trajectory (i.e., main path)
   - There are various methods

Assigning weight



- Search Path Link Count (SPLC) algorithm
  - Weight of 5→7=25, as each of five preceding nodes (1 to 5) traverses the arc five times to reach the sink nodes (9 to 11)
  - Weight of 8→12=12, as four of its preceding nodes (1 to 4) traverses the arc twice (one following the arc 6→8 and the other following the arc 5→7), and the other four preceding nodes (5 to 8) traverse the arc once, to reach the sink node 12

Determining main path



- Global Search method
  - Selects one of the paths from source to sink nodes having the greatest combined weight
  - The main path in the network involves the source nodes 1 to 3, the intermediate nodes 4, 5, 7, 8, and the sink nodes 12, 13, and the combined weight along the path is 76 (=7+20+25+12+12)

Outlines

- Background
- ➢ **Methodology**
- Empirical Result
- Conclusion

Arc's structural connectivity

- An arc would have a greater weight if the arc has greater structural connectivity
  - meaning that it can be reached from more preceding nodes and/or it may lead to more succeeding nodes
- The arc may be deemed as more important in disseminating knowledge

Drawbacks of MPA

- All arcs are treated equal
  - Regardless of how far away the arc is from the source
    - Arc weight should "decay" down the citation chain
  - Regardless of the degree of similarity between the cited and the citing ones
    - Two ideas
      - Higher weight for more similar pairs
      - Higher weight for less similar pairs

Modified analysis procedure

1. A citation network is constructed
2. A weight for each arc is assigned based on its (traversal count) x (similarity between cited and citing's sets of patent classification symbols)
3. A series of connected arcs across the network is determined as a representative trajectory (i.e., main path)

Patent classification

Current U.S. Class:	**715/863**; 345/173; 345/179
Current CPC Class:	G06F 3/04883 (20130101); G06F 21/36 (20130101); H04M 1/663 (20130101); G06F 3/0488 (20130101); G06F 3/017 (20130101); G06F 3/0484 (20130101); G06F 3/04842 (20130101); H04M 1/67 (20130101); H04M 1/575 (20130101); H04M 2250/22 (20130101)
Current International Class:	G06F 3/033 (20060101)

- Every patent is assigned **one or more classification symbols** during its application process by examiner according to the patent's disclosed invention and **a standard scheme** such as IPC (international patent classification), CPC (cooperative patent classification), etc.

Patent classification

- Patents' classification symbols are a valuable source of information
  - determined by **professional examiner**
  - representative of the patents' **technical contents**
  - based on **a common standard**

Levels of symbols

- Classification symbols are usually reduced to a higher level for easier processing
  - CPC 3rd level symbols ≒ 700
  - CPC 4th level symbols ≒ 8000

3989811

CPC	3rd level	4th level
B01D 53/1456	B01D	B01D 53/00
B01D 53/526	B01D	B01D 53/00
C10L 3/104	C10L	C10L 3/00
C10L 3/103	C10L	C10L 3/00
C01B 17/0408	C01B	C01B 17/00

4336233

CPC	3rd level	4th level
B01D 53/1493	B01D	B01D 53/00
C10K 1/14	C10K	C10K 1/00
C10K 1/143	C10K	C10K 1/00
Y02P 20/152	Y02P	Y02P 20/00
Y02C 10/06	Y02C	Y02C 10/00

Similarity between symbols

- Jaccard coefficient= $J(x,y)=|x\cap y|/|x\cup y|$
- Generalized Jaccard coefficient

$$J(x,y)=\frac{\sum_i \min(x_i,\ y_i)}{\sum_i \max(x_i,\ y_i)},\qquad 1\le i\le n.$$

- Example: patents 3989811 and 4336233
  - Jaccard=1/6, Generalized Jaccard =1/9

	B01D	C10L	C01B	C10K	Y02C	Y02P
3989811	2	2	1	0	0	0
4336233	1	0	0	2	1	1
$\min(x_i, y_i)$	1	0	0	0	0	0
$\max(x_i, y_i)$	2	2	1	2	1	1

Outlines

- Background
- Methodology
- ➢ **Empirical Result**
- Conclusion

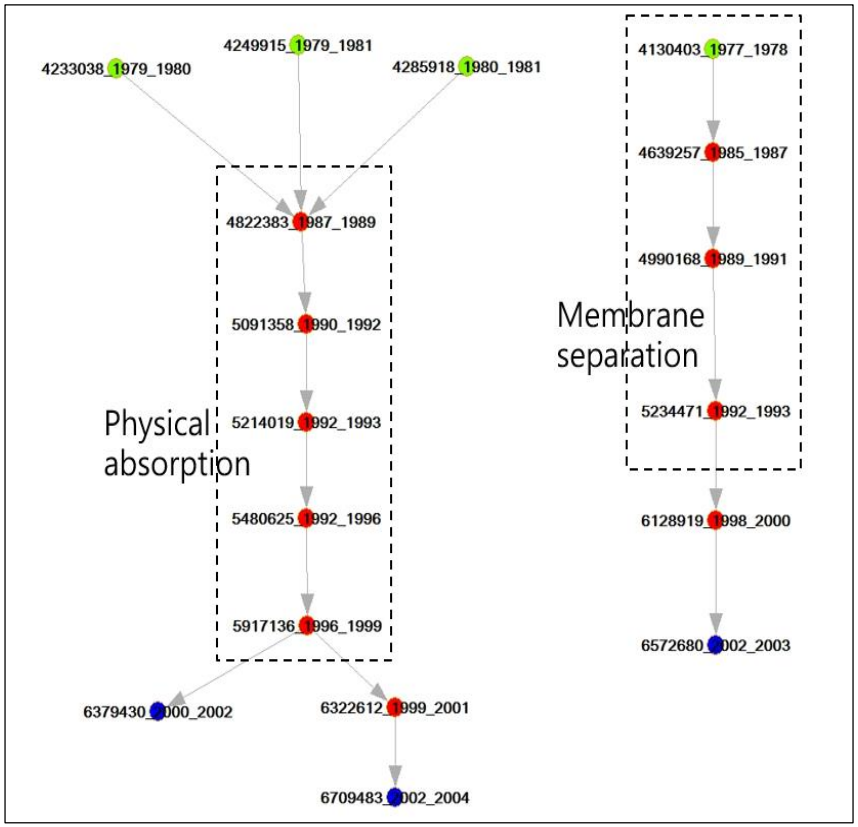
Empirical data

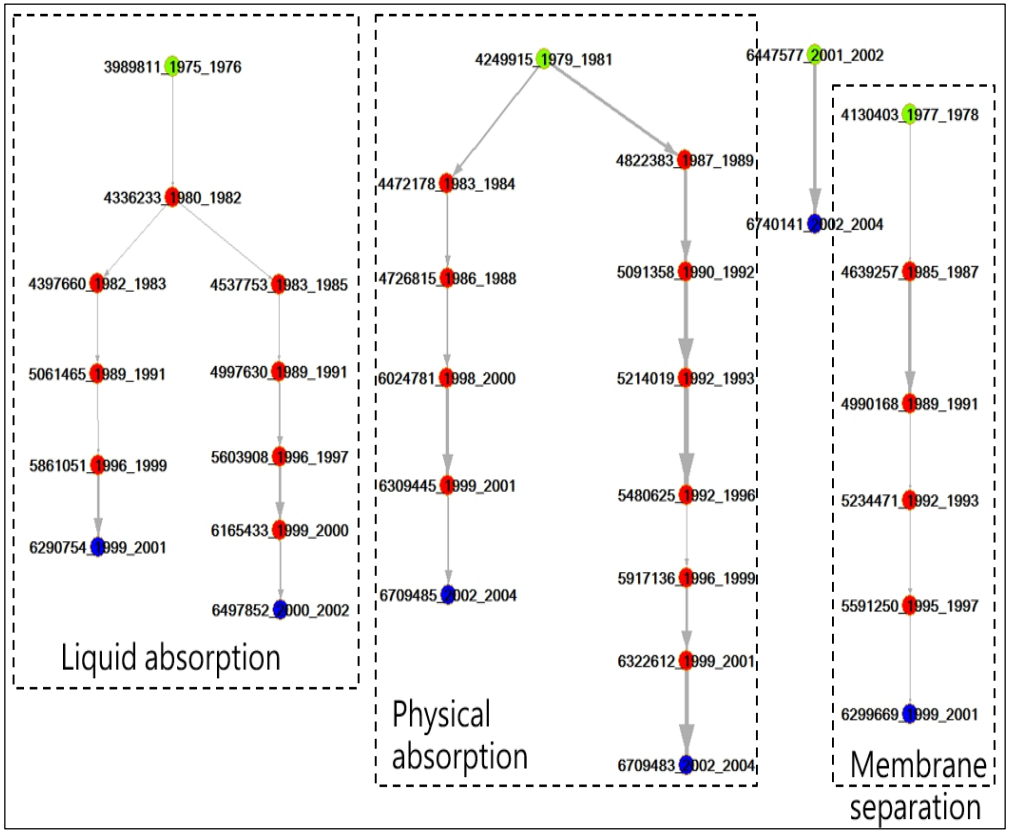
- 675 USPTO patents related to CO_2 capture and storage
- Using CPC symbols
  - Observing difference between 3rd level and 4th-level symbols
- Jaccard similarity
  - Observing difference between ordinary and generalized Jaccard

Observation

- Using 4th level symbol reveals more consistent lineage of technology development
- Using generalized Jaccard reveals more consistent lineage of technology development

An Example





Outlines

- Background
- Methodology
- Empirical Result
- **Conclusion**

Conclusion

- Pros
  - A main path better reflects the lineage of technology development
  - The modified MPA may also identify additional trajectories of technology development
- Cons
  - For a large citation network, the traversal counts of the arcs become so great that the limited Jaccard coefficients cannot provide much differentiation

Thank You