Missing Links:
Timing characteristics and their application on capturing contemporaneous technology developments

Kuan, Chung-Huei
Assistant Professor
Graduate Institute of Patent
National Taiwan University of Science and Technology

Chen, Dar-Zen
Professor
Dept. of Mechanical Engineering &
Inst. of Industrial Engineering
National Taiwan University

Outline

- *Introduction*
- Data
- Methodology
- Analysis and result
- Conclusion

Introduction

- Tools for capturing patent relatedness
  - Citation-based
    - Direct citation
    - Bibliographic coupling
    - Co-citation
  - Text-based
  - Classification-based

Focus of this paper

Introduction

- What is a missing link (*ML*)?
  - A *missing link* occurs between two patents *E* and *L* if
    - *E* and *L* do not directly cite each other;
    - *E* and *L* are bibliographic coupled; and
      - *E* and *L* have high bibliographic coupling strength (BCS)

Introduction

- A ML example
  - US utility patents, US8,622,222 (*E*) and US8,623,202 (*L*)
    - Both concern membrane bioreactor technologies
      - Both filed by the same company,
        - one in January 2011 and the other in October 2012
      - Both granted in January 2014 by different examiners
    - The two patents do not cite each other but have exceptionally high BCS of 1,039
      - US8,622,222 cited 1,063
      - US8,623,202 cited 1,072
        - domestic and foreign patents and published applications

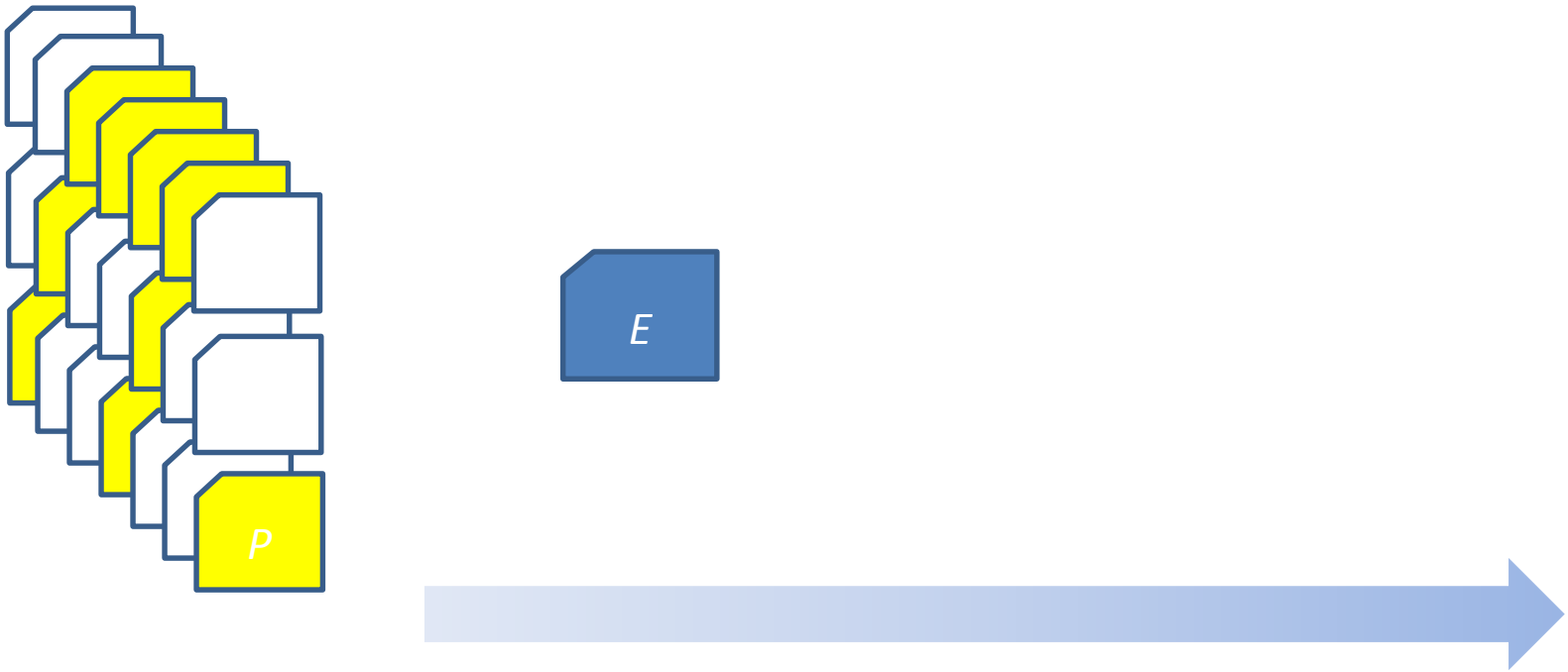
Introduction

- What is a missing link (*ML*)?
  - A *conflicting phenomenon* between direction citation (DC) and bibliographic coupling (BC)
    - *E* and *L* do not directly cite each other
    - *E* and *L* are strongly bibliographic coupled
    - Do *E* and *L* related or not?

Introduction

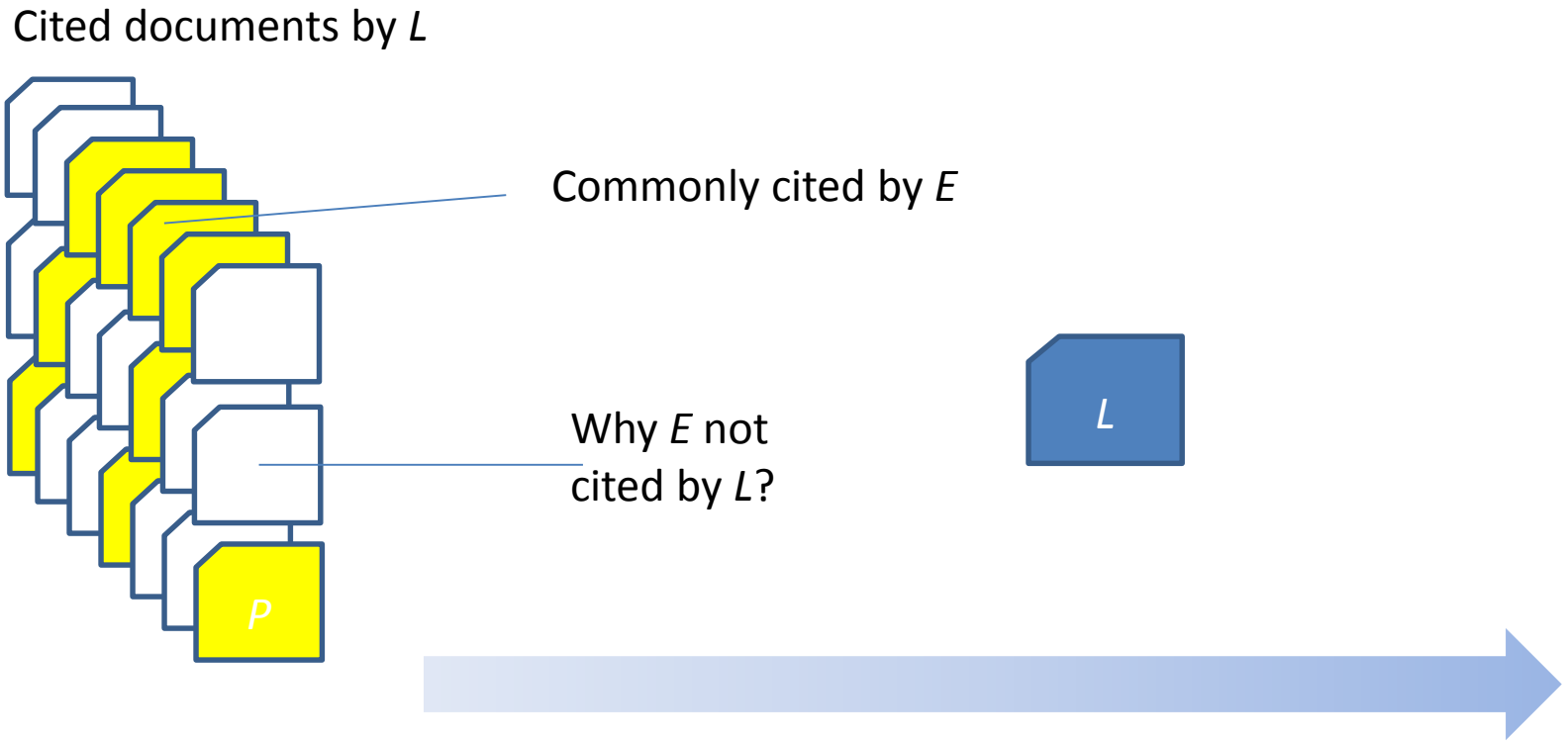
- What is a missing link (*ML*)?
  - A *conflicting phenomenon* between direction citation (DC) and bibliographic coupling (BC)

Cited documents by *E*



Introduction

- What is a missing link (*ML*)?
  - A *conflicting phenomenon* between direction citation (DC) and bibliographic coupling (BC)



Introduction

- ML may be utilized to discover patent relatedness that escapes detection using DC
- This study addresses the following issues
  - Why MLs occur and whether they are simply coincidences
  - What useful information may be captured by MLs if their occurrence is not coincidental
  - How MLs may be utilized to capture this useful information

Outline

- Introduction
- *Data*
- Methodology
- Analysis and result
- Conclusion

Data

- Patents in the field of carbon dioxide capture, storage, recovery, delivery, and regeneration and collects
- A total of 34,083 US utility patents issued between 1976/1/1 and 2017/3/31
- Among the 34,083 patents, there are 155,076 DC and 1,609,549 BC pairs.
  - The BC pairs have a significantly skewed BCS distribution
    - A mean (μ) of 2.74, a standard deviation (σ) of 15.66, and a maximum of 1,123
    - 72.55% (1,167,794) have the smallest BCS of 1
  - 75,700 are both BC and DC pairs
    - much higher mean BCS (9.56) than the overall average

Data

- 9,213 ML pairs are identified
  - Each pair does not cite each other
  - Each pair has BCS greater than 34 (=μ + 2σ)
- BCS distribution for the 9,213 ML pairs

Range	35-100	101-200	201-300	301-400	401-500	501-600	601-700	701-800	801-900	901-1,000	1,001-1,200
Pairs	5,686	1,987	1,001	221	139	54	38	32	27	25	3
%	61.72	21.57	10.87	2.40	1.51	0.59	0.41	0.35	0.29	0.27	0.03

More than 93%

Data

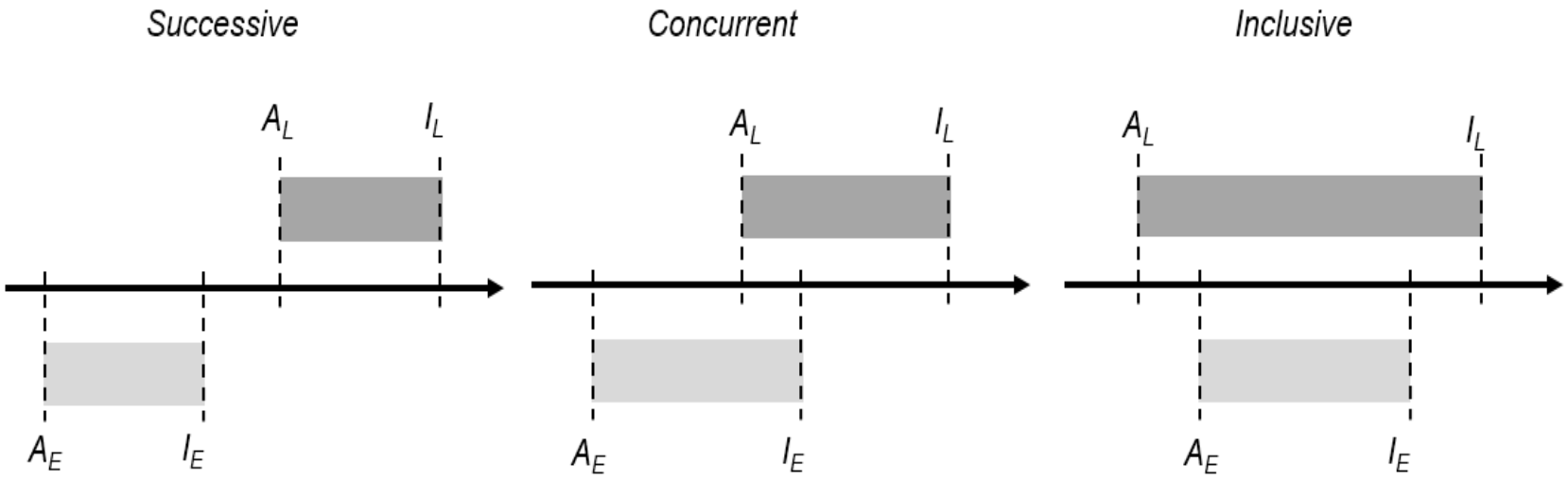
- A fixed BCS threshold may not be appropriate
  - It may exclude some critical BC pairs in an early stage of technology evolution
  - It may include less important ML pairs in later stages
  - However it enables efficient observations as different amounts of MLs are supplemented

Outline

- Introduction
- Data
- *Methodology*
  - *Timing characteristics of DCs/MLs*
  - *Main path analysis*
- Analysis and result
- Conclusion

Methodology

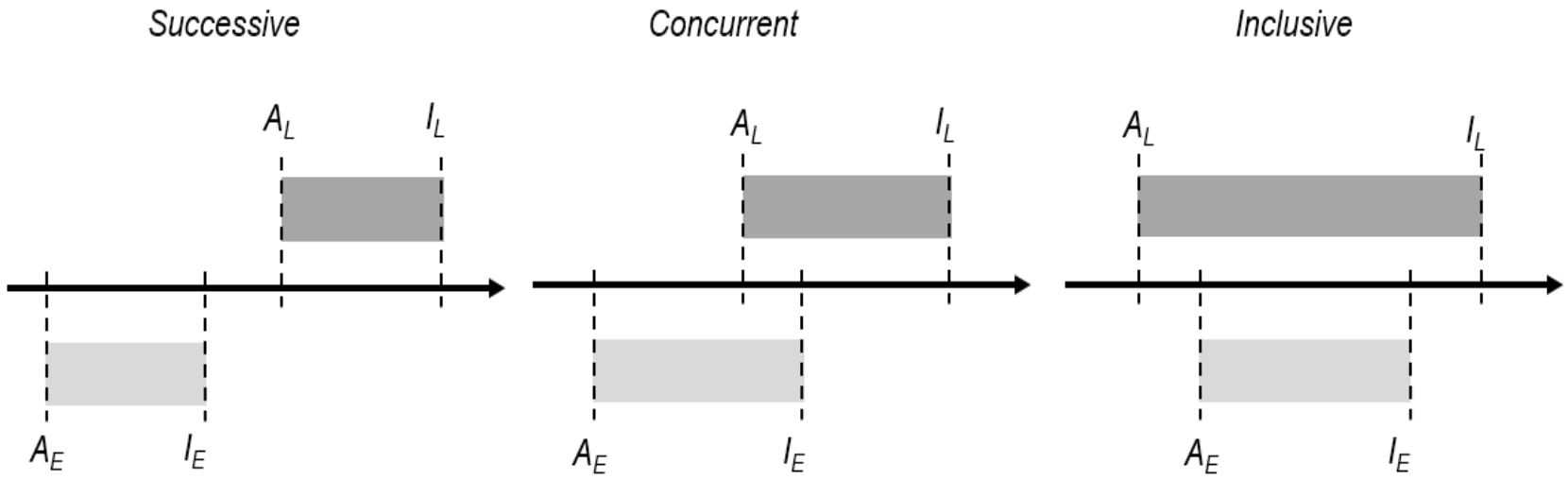
- Timing Characteristics of DCs/MLs



- Application dates are denoted as A_E and A_L and issue dates as I_E and I_L, with subscripts *E* and *L* representing the pair's *earlier* and *later* issued patents,

Methodology

- Timing Characteristics of DCs/MLs



	Successive		Concurrent		Inclusive	
Timing characteristics	$A_E < I_E < A_L < I_L$		$A_E \leq A_L \leq I_E \leq I_L$		$A_L \leq A_E < I_E \leq I_L$	
Later applicant citing earlier	Possible		Unlikely		Unlikely	
Later examiner citing earlier	Possible		Possible but rare		Unlikely	
ML pairs	2,214	24.03%	4,196	45.54%	2,803	30.42%
DC pairs	169,916	90.04%	16,003	8.48%	2,802	1.48%

Methodology

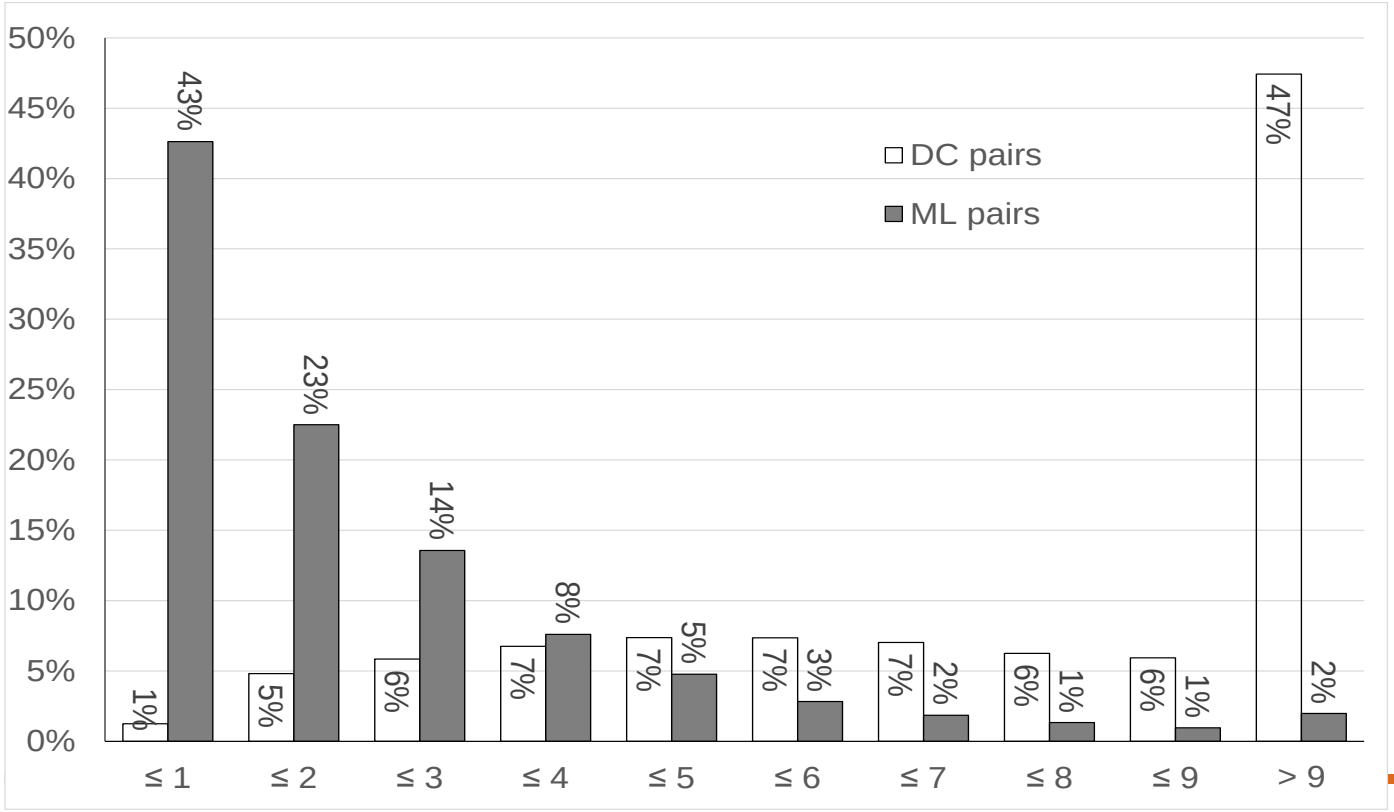
- Timing Characteristics of DCs/MLs

	Successive		Concurrent		Inclusive	
Timing characteristics	$A_E < I_E < A_L < I_L$		$A_E \leq A_L \leq I_E \leq I_L$		$A_L \leq A_E < I_E \leq I_L$	
Later applicant citing earlier	Possible		Unlikely		Unlikely	
Later examiner citing earlier	Possible		Possible but rare		Unlikely	
ML pairs	2,214	24.03%	4,196	45.54%	2,803	30.42%
DC pairs	169,916	90.04%	16,003	8.48%	2,802	1.48%

- ML pairs
  - Three-quarters (75.97% = 45.54% + 30.42%) belong to the *concurrent* and *inclusive* types
- DC pairs
  - Only approximately 10% (9.96% = 8.48% + 1.48%) belong to the *concurrent* and *inclusive* types

Methodology

- Timing Characteristics of DCs/MLs
  - Distributions of time spans, $I_L - I_E$ in years, of the DC and ML pairs



Methodology

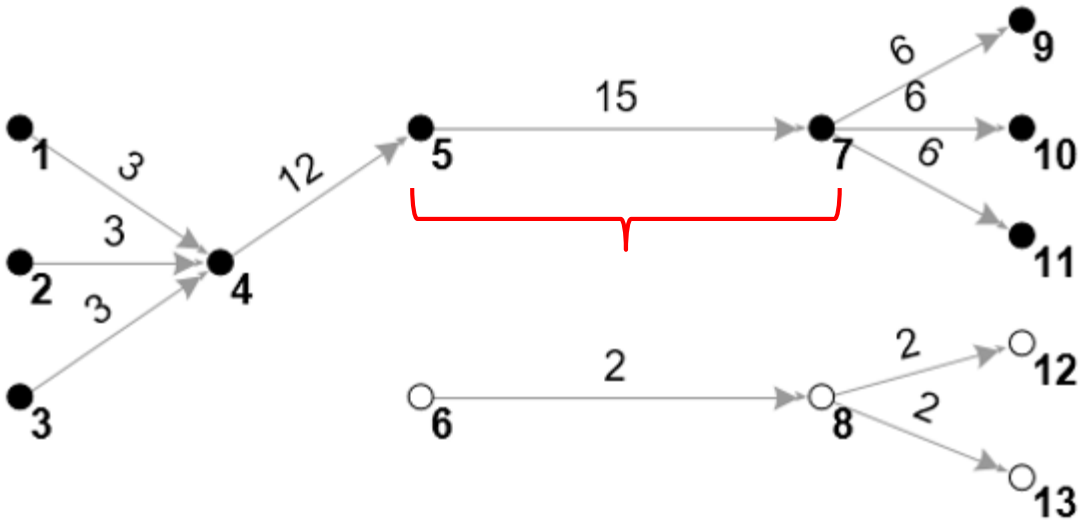
- Timing Characteristics of DCs/MLs
  - ML pairs
    - usually have highly overlapped application processes
  - DC pairs
    - more frequently have successive or less overlapped application processes
  - MLs are not coincidences and may identify patent relatedness when DC is less likely.

Methodology

- Main path analysis (MPA)
  - Determines the major development trajectory of a scientific field by identifying the most significant chains of DCs in a citation network of scientific articles
  - Generally involves three major steps
    - A citation network is constructed
    - A weight for each arc is determined according to the arc's traversal count
    - A series of connected arcs across the network is determined as a representative trajectory, referred to as the *main path* of the citation network.

Methodology

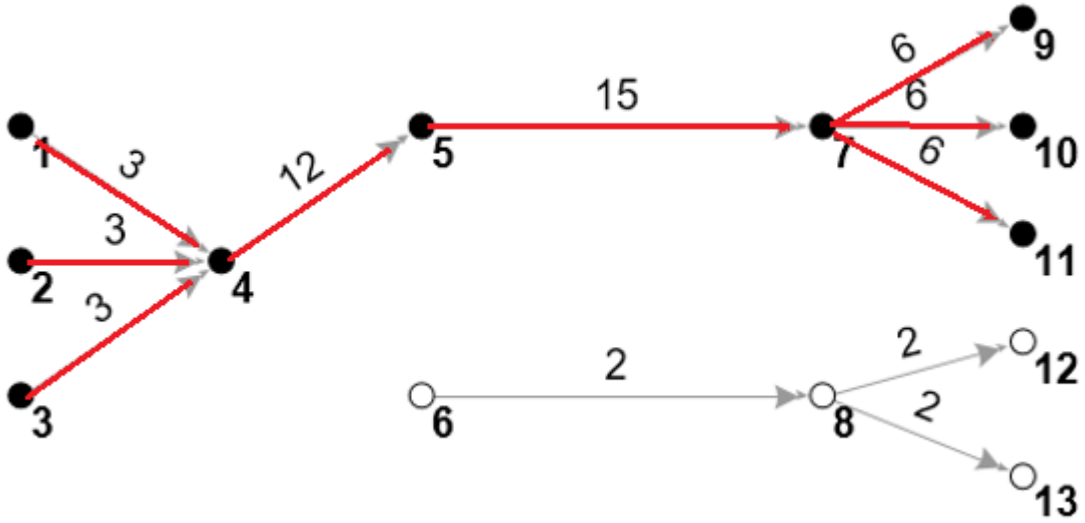
- Main path analysis (MPA)



  - There are various algorithms for weights
  - Using the algorithm *search path link count* (SPLC), the weight of the arc 5→7 is 15
    - SPLC counts the number of traversals of the arc 5→7 from all preceding nodes (1 to 5) to the sink nodes (9 to 11)
  - No matter the algorithm, an arc has greater weight if it has greater structural connectivity

Methodology

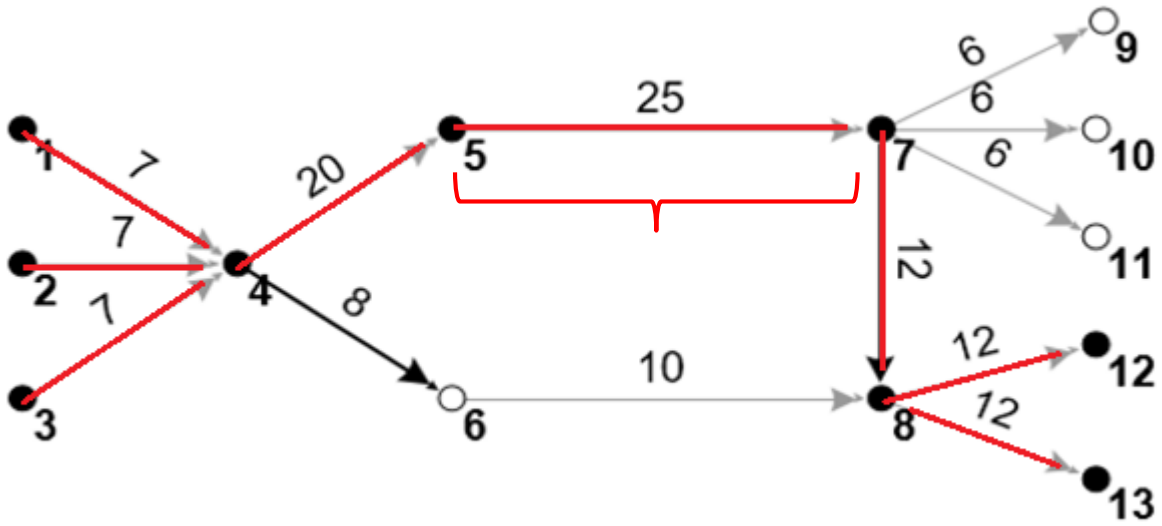
- Main path analysis (MPA)



- Various methods also exist for determining the main path.
  - *Global search* selects the path from source to sink nodes having the greatest total weight.
  - *Local search* begins from the source nodes, selects the arc(s) from these nodes with the greatest weight(s), and works forward for the next search until a sink node is reached
  - *Key-route* determines one or more main paths by locating the arc(s) having the greatest weight first and tracing both backward and forward until source and sink nodes are reached.
- The main path determined using global search method are the one connecting the black nodes.
  - Total weight = 3+12+15+6=36

Methodology

- Main path analysis (MPA)



- Using the algorithm *search path link count* (SPLC), the weight of the arc 5→7 is 25
  - There are five preceding nodes (1 to 5) and each will traverse the link 5→7 once to reach one of the sink nodes (9 to 13)
- The main path determined using global search method are the one connecting the black nodes
  - Total weight =7+20+25+12+12=76

Methodology

- Main path analysis
  - This study extends MPA to a network involving not only DCs but also MLs (therefore, a *heterogeneous* network)
    - Both the *explicit* relatedness manifested by the DCs and the *latent* relatedness captured by the MLs are considered
    - To incorporate MLs into the network, each ML pair is represented by an arc originating from the pair's lower-numbered patent to the pair's higher-numbered patent

Outline

- Introduction
- Data
- Methodology
- *Analysis and result*
- Conclusion

Analysis and Result

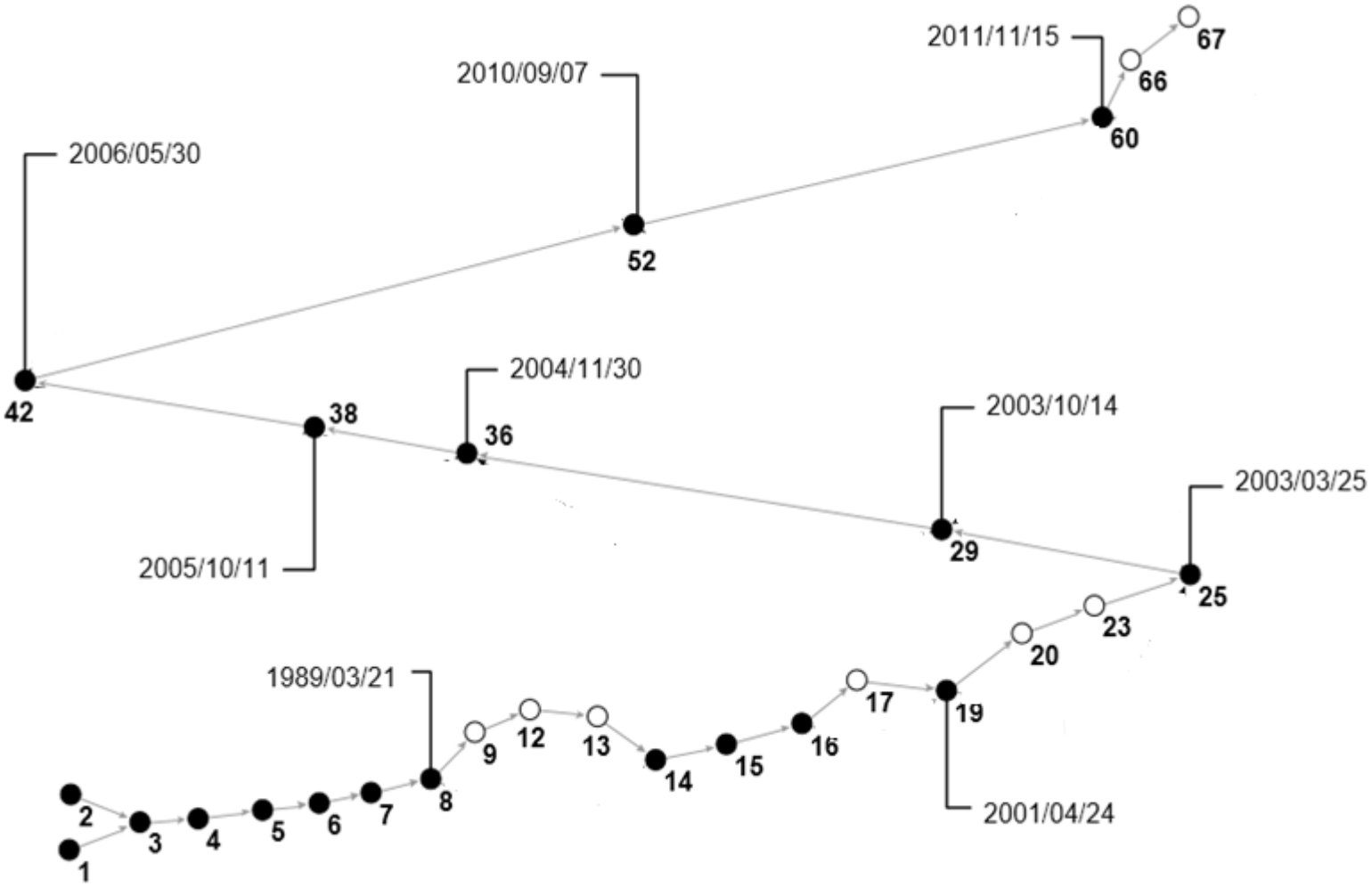
- Two networks are constructed using the empirical data
  - A conventional PCN with 34,083 nodes connected by 155,076 arcs, one for each DC pair
  - A heterogeneous network with an additional 9,213 arcs, one for each ML pair
- Arc weights are assigned using the *SPLC* algorithm
- Main paths are derived using the *global search* method
  - The path obtained from the PCN is referred to as the *original main path* (OMP)
  - The one obtained from the heterogeneous network is referred to as the *heterogeneous main path* (HMP)

Analysis and Result

- Black nodes: patents by both OMP and HMP
- White nodes: patents by the OMP only
- Gray nodes: patents by the HMP only
- Solid gray arc: DC
- Dashed black arc: ML

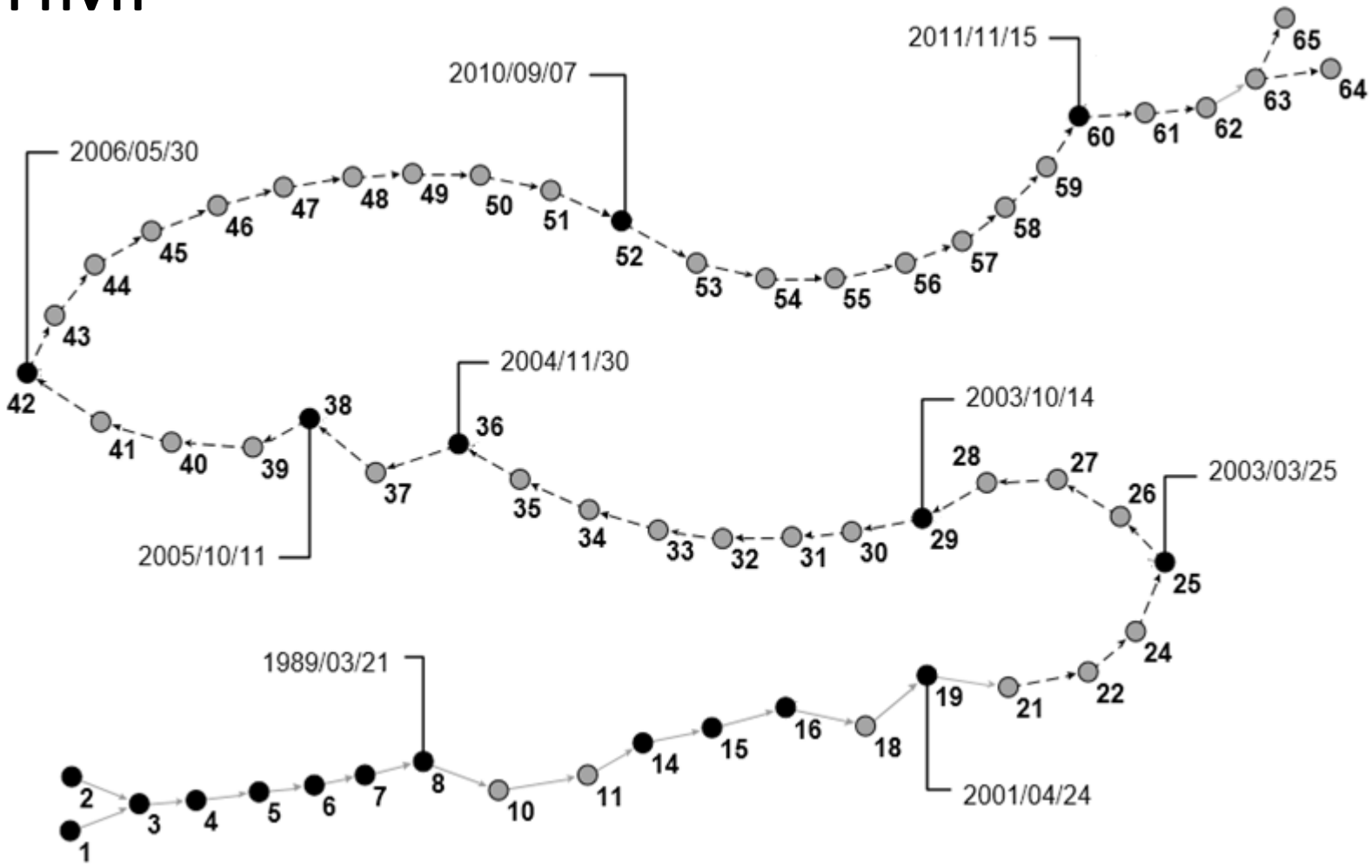
Analysis and Result

- OMP

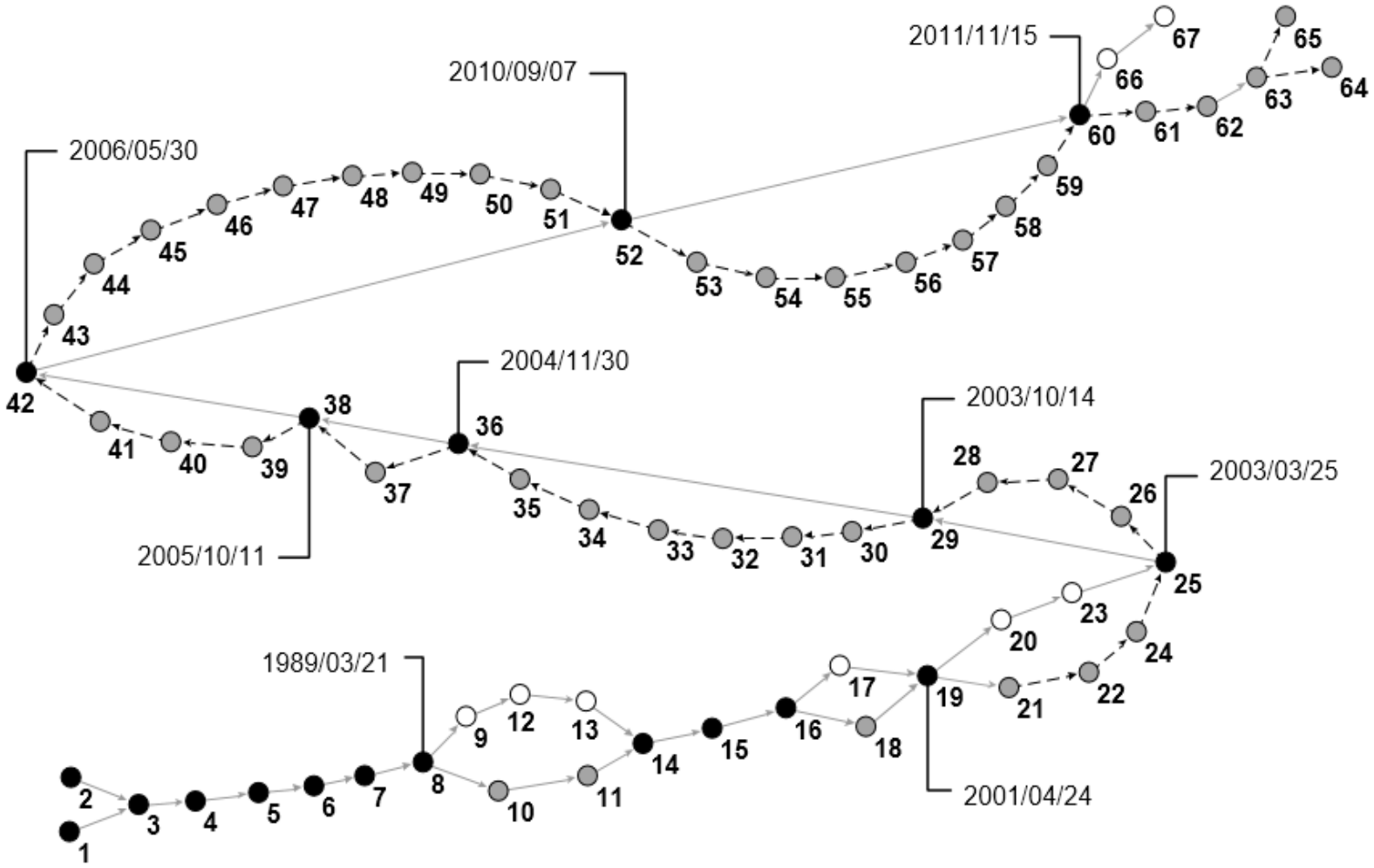


Analysis and Result

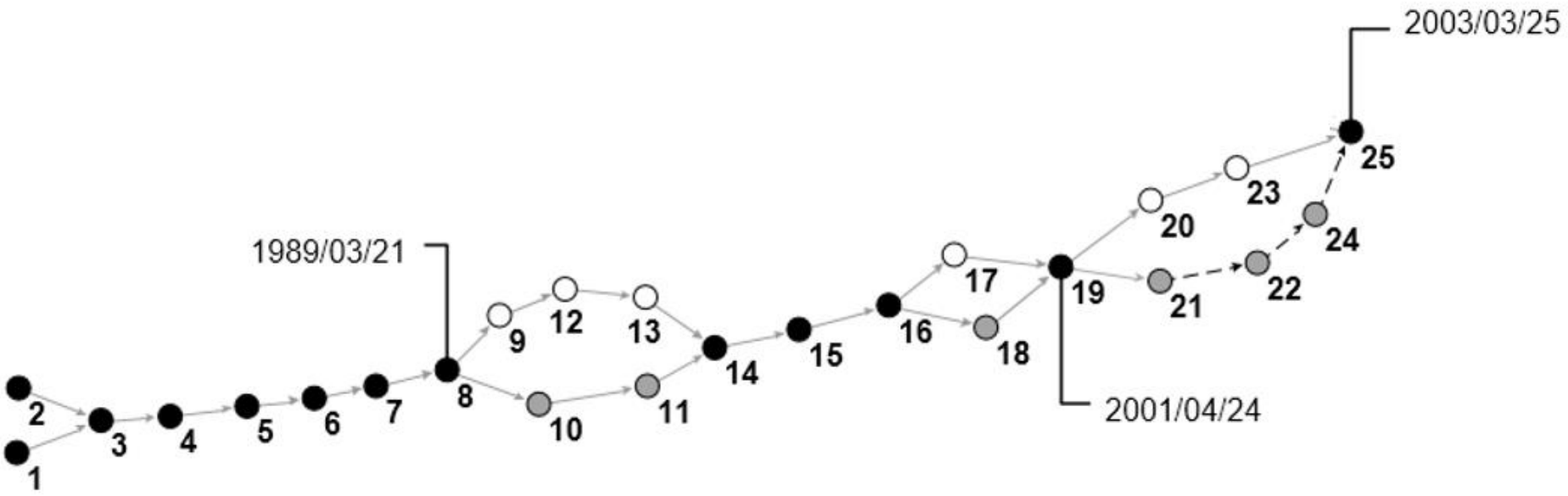
- HMP



Analysis and Result

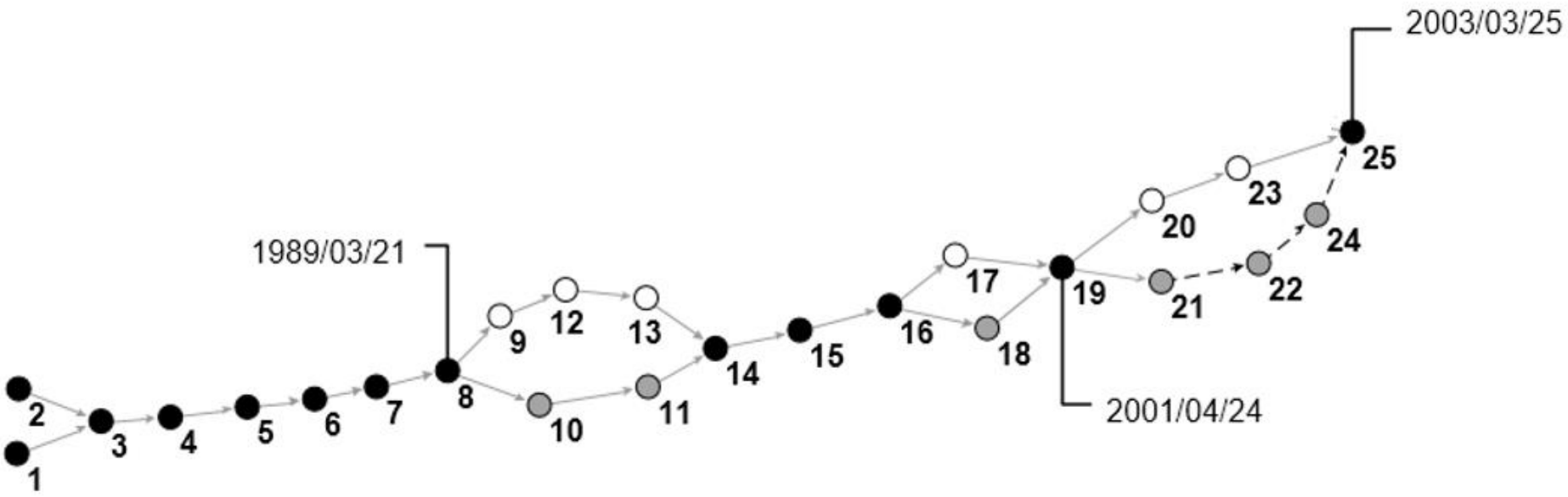


Analysis and Result



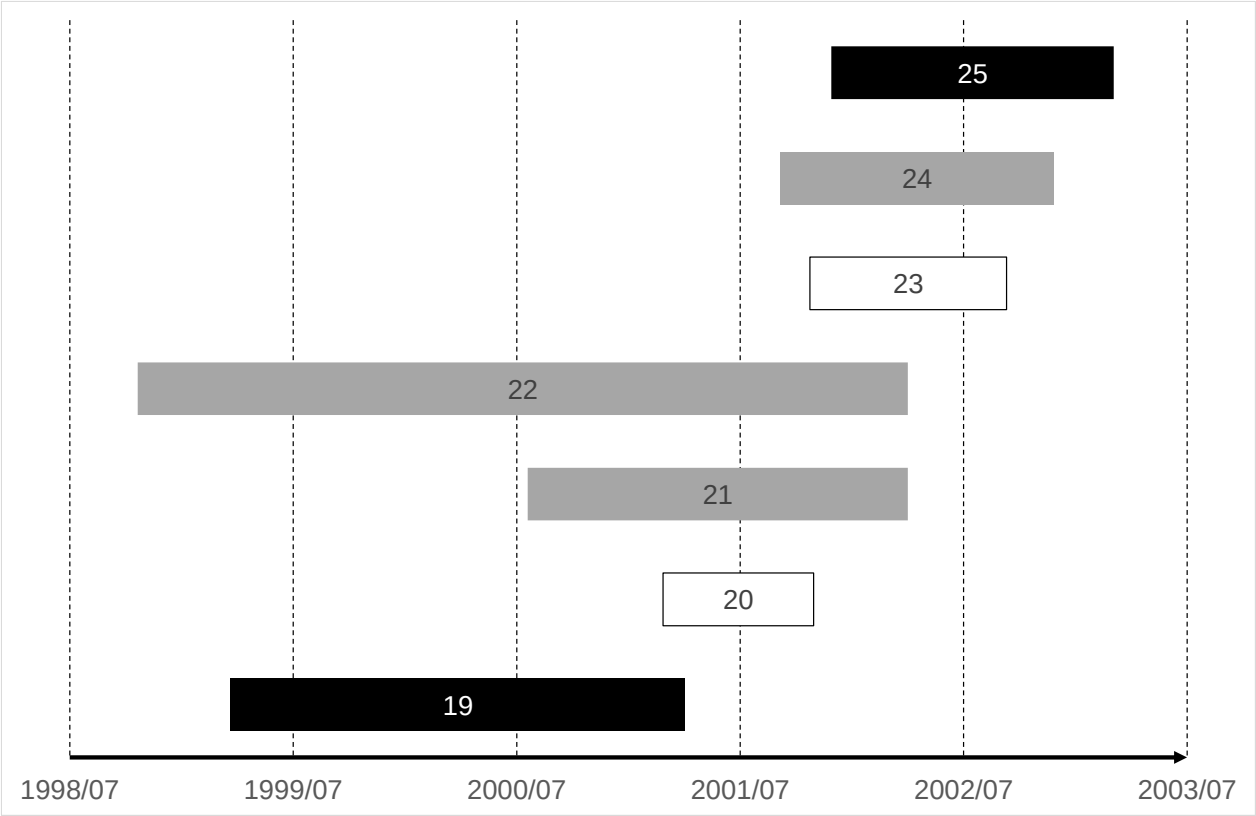
#	Time window	HMP seg.	MLs	DCs	MLs/DCs	Avg. Doc. Cited
1	1976/01/01~1989/03/21	1→…→8	11	5,885	0.19%	8.00
2	1989/03/22~2001/04/24	8→…→19	1,513	118,005	1.28%	15.69
3	2001/04/25~2003/03/25	19→…→25	985	19,472	5.06%	20.02

Analysis and Result



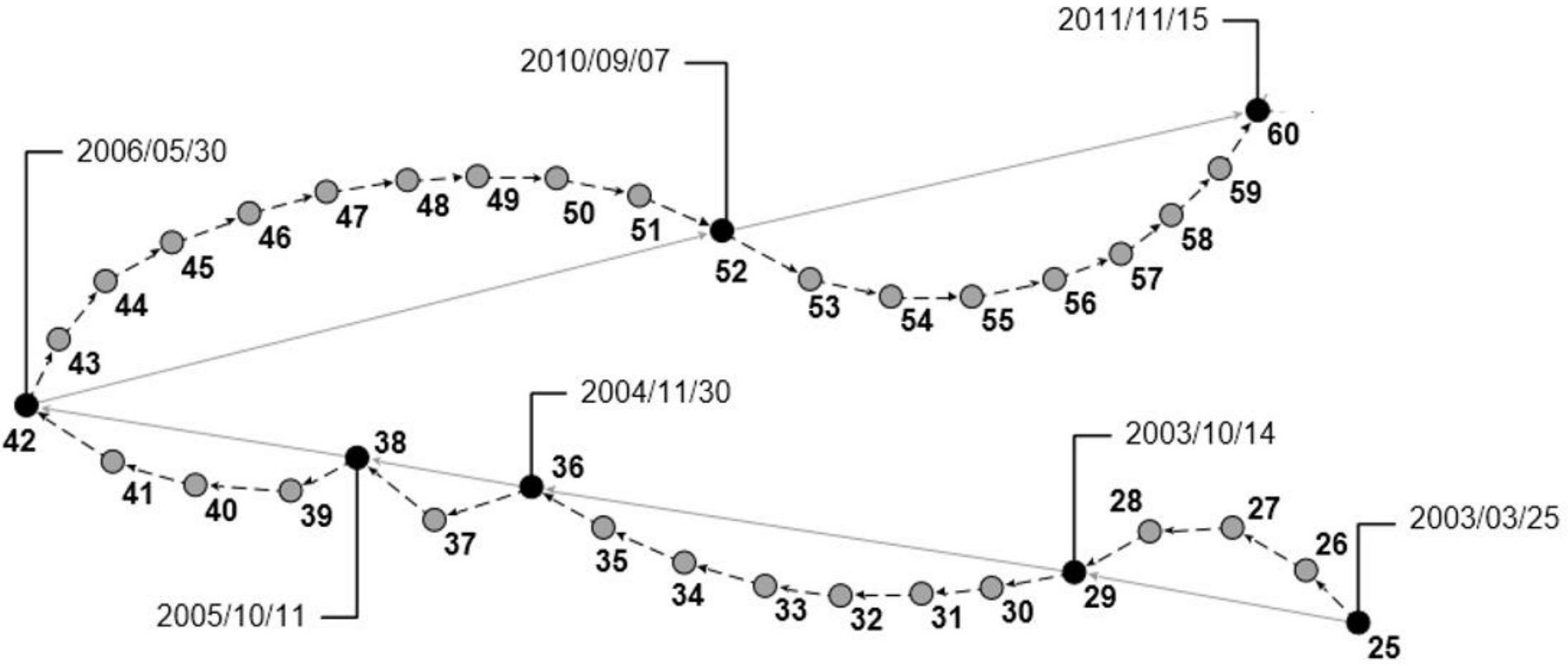
#	Time window	HMP seg.	MLs	DCs	MLs/DCs	Avg. Doc. Cited
1	1976/01/01~1989/03/21	1→…→8	11	5,885	0.19%	8.00
2	1989/03/22~2001/04/24	8→…→19	1,513	118,005	1.28%	15.69
3	2001/04/25~2003/03/25	19→…→25	985	19,472	5.06%	20.02

Analysis and Result



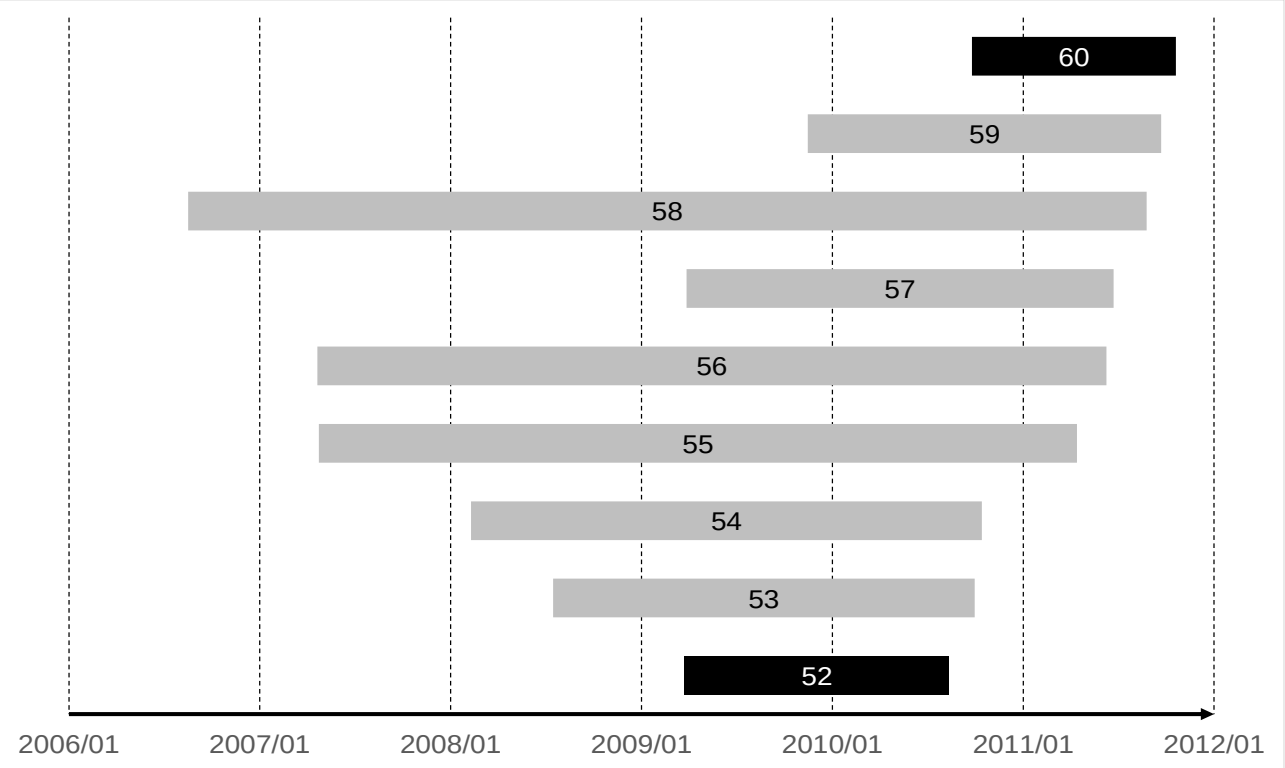
- The application processes of patents in the original route (i.e., nodes 19, 20, 23, and 25, connected by DC arcs) overlap less than those of patents in the new route (i.e., nodes 19, 21, 22, 24, and 25

Analysis and Result



#	Time window	HMP seg.	MLs	DCs	MLs/DCs	Avg. Doc. Cited
4	2003/03/26~2011/11/15	25→…→60	5,200	42,594	12.21%	30.52

Analysis and Result



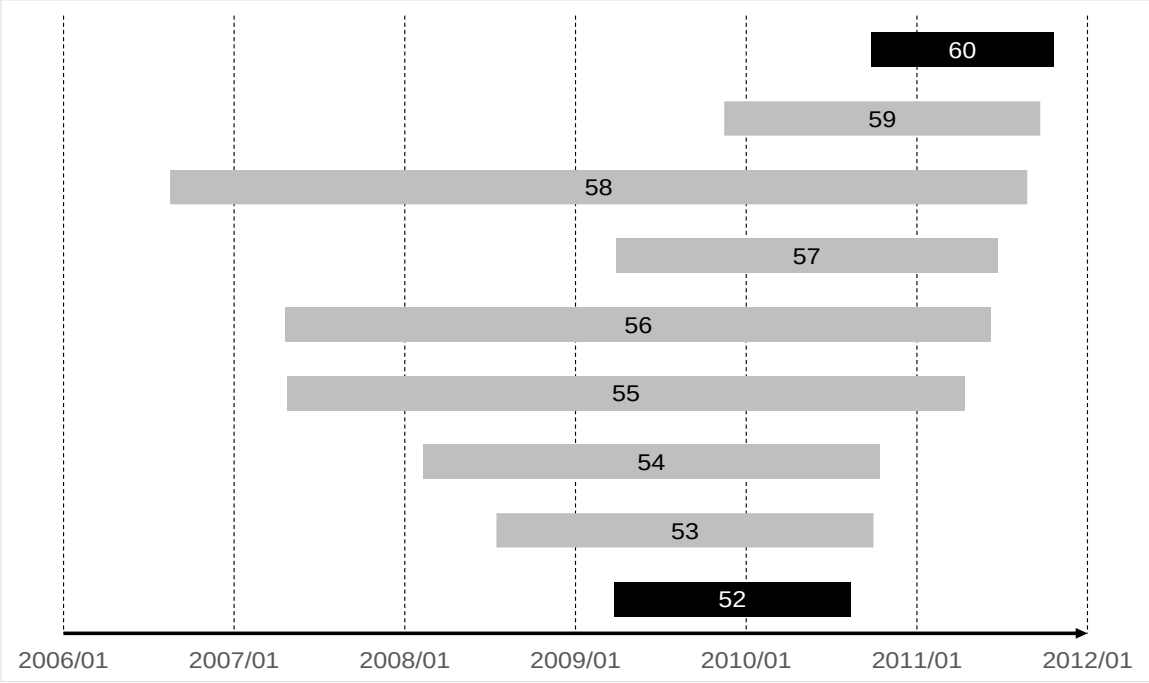
- Most patents within this subwindow have an inclusive relationship with an earlier patent and/or a later patent
  - For example, the application process of patent 52 was covered by that of patent 53, whose application process was covered by that of patent 54, and so on

Analysis and Result

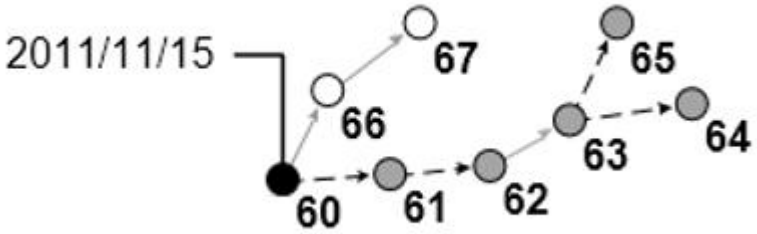
- These patents embody contemporaneous technological developments
  - As technology evolves, developments of related technologies occur concurrently
  - If these contemporaneous developments are filed for patent protection, their application processes mostly overlap
  - Their relatedness then may elude detection by DCs but would still be caught by MLs
- MLs therefore appear to fill the “gap” between a DC pair by identifying these highly related patents that reflect contemporaneous developments.

Analysis and Result

- The HMP introduces a sequential order among the contemporaneous patents
  - This sequential order should not be taken at face value



Analysis and Result



#	Time window	HMP seg.	MLs	DCs	MLs/DCs	Avg. Doc. Cited
5	2011/11/16~2017/03/31	60→…→67	1,504	2,766	54.37%	40.83

- Patents issued at this late stage had more documents available for citation, and patent pairs are more likely to pass the BCS threshold (34) and qualify as ML pairs
- The numerous MLs causes HMP to extend independently from the OMP

Analysis and Result

- The sporadic ML arcs of windows 1 and 2 have little influence
  - The resulting HMP does not differ significantly from the OMP
- As the density of ML arcs increases in windows 3 and 4
  - These MLs fill the gaps left open by DCs
  - These ML arcs join to form new routes with stronger structural connectivity
  - A more comprehensive trajectory is obtained
- When the density of ML arcs continues to rise, the HMP may deviate entirely from the OMP

Outline

- Introduction
- Data
- Methodology
- Analysis and result
- *Conclusion*

Conclusion

- This study finds that
  - ML pairs often undergo highly overlapped application processes and the applicants/examiners of such ML pairs are inherently unlikely to cite each other
    - DC pairs more frequently have successive or less overlapped application processes, and their applicants/examiners are not handicapped in citing each other
  - MLs are fostered out of systematic context
  - MLs may capture patent relatedness that DCs fail to detect.
    - Analysts should not omit MLs when conducting patent citation analysis; otherwise, some crucial patent relatedness may be systematically ignored.

Conclusion

- This study proposes a method of utilizing MLs for the investigation of technological development by extending MPA to a heterogeneous network
- This study finds that
  - MLs capture concurrent efforts in developing related technologies embodied in contemporaneous patents
  - By identifying these contemporaneous patents, analysts should be able to acquire a more complete and thorough understanding of the evolution of technology

Thank you