Chapter 8
Patterns and Trends in Semantic Predications

Abstract We demonstrate a series of studies of semantic predications from Semantic MEDLINE, including the detection of semantic predications with burstness and in association with conflict, contradictory, or other sources of uncertainties of scientific knowledge. Semantic networks of predications are analyzed within the framework of structural variations. Examples in this chapter represent scientific knowledge at a level of granularity that differs from those studies of scientific knowledge at the level of articles or journals of scholarly communication.

Semantic MEDLINE Database

The backend of Semantic MEDLINE is the Semantic MEDLINE Database (SemMedDB) (Kilicoglu et al. 2012). As of December 31, 2016, SemMedDB contains about 89.2 million predications from 26.7 million bibliographic records from MEDLINE. Its primary coverage is the biomedical literature. These predications are extracted by SemRep. The current version of SemMedDB is semmever30.

Representing Semantic Predications as a Graph

SemMedDB contains several tables of citations (in the MEDLINE sense of the term), i.e. the metadata of a published article, original sentences, and predications. For example, the SENTENCE table contains information on individual sentences such as SENTENCE_ID, PubMed ID (PMID), and the sentence. The PREDICATION table contains various information about predications such as PREDICATION_ID, a SENTENCE_ID, PMID (PubMed ID), PREDICATE, SUBJECT_CUI, SUBJECT_NAME (preferred name of the subject of the predication), and similar fields for the object of the predication. We loaded SemMedDB version 24 to a MySQL database. The examples explained below are based on this version. Figure 8.1 shows a

C. Chen and M. Song, *Representing Scientific Knowledge*,
https://doi.org/10.1007/978-3-319-62543-0_8

visualization of a network of semantic predications in Neo4j, a graph database. The visualization shows that the semantic connections are unevenly distributed. Some entities are connected by a lot of semantic relations, whereas some are connected by few connections. The unevenness implies a level of uncertainty.

A distinct advantage of a graph database over the traditional relational database is a reduced complexity of queries. As illustrated in Table 8.1, a complex and time-consuming query with multiple table joins in a relational database can be reduced to a simple and efficient query in a graph database in Neo4j with the Cypher query language. The query in the graph database is in Cypher, a powerful query language supported by Neo4j. The query is to find paths that start with a doctor node and connect to a therapy node through at least four other types of nodes in between. A Cypher query shares some similarities with MySQL queries in terms of their style.

A study of scientific claims often need to address a series of questions based on the current results of a search. Graph databases such as Neo4j provide the desirable flexibility. Consider the following questions concerning the interest in scientific knowledge relevant to a body of scientific publications. These questions may remind you Heilmeier's Catechism we discussed in Chap. 1.

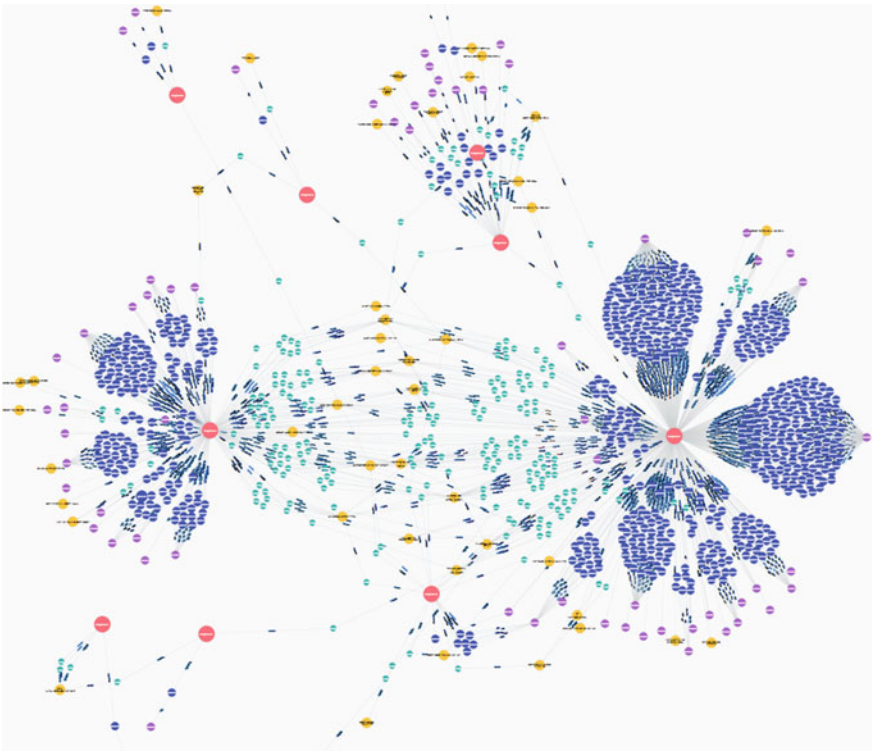

Fig. 8.1 A network of semantic predications visualized in Neo4j

Table 8.1 The complexity of a query can be reduced in a graph database

Database Type	Query
Relational	SELECT C.DiagnosisID as diagnosis, C.TherapyID as therapy FROM prescriptions as C JOIN (SELECT DiagnosisID as ID, A.ICD10_CODE as code, A.ICD10_TEXT as text FROM diagnosis as A JOIN (SELECT PracticeID, DoctorID, PatientID, Age, DiagnosisID, Action FROM actions ON A.DiagnosisID=B.DiagnosisID) as D ON C.DiagnosisID=D.ID;
Graph	MATCH path=(d:DOCTOR)-[r*4..]- (t:THERAPHY) RETURN path LIMIT 100;

1. When did a claim (hypothesis, assertion, conjecture, statement) appear for the first time in a body of source information, e.g. the literature? In which paper/source?
2. How certain was the assertion?
3. How many papers made the same claim subsequently?
4. How many papers made claims that contradict a given claim subsequently?
5. What are the closely related but distinct claims for a given claim?
6. Given a claim, which reference is most frequently cited along with the claim's citation context?
7. Which references are frequently co-cited in relation to a given claim?
8. How many dimensions (clusters or eigenvectors or topics) are associated with the citation contexts of a claim?

There are several advantages of addressing these questions in a graph database. In particular,

1. Much faster responses than using relational databases such as MySQL
2. Much more flexible to formulate complex queries for complex questions
3. Much easier to incrementally update the database
4. Particularly suitable for detecting emerging trends in research.

We illustrate the flexibility of the generic approach with an example of 13 full-text publications of our own. The small graph contains 12 authors, 48 cited references, and 36 sentences that contain citations. The approach is applicable to a wide variety of subject areas regardless their overall uncertainty levels because the mechanisms for differentiating uncertainties from claims will be in place as a unique feature of the approach. The Cypher query below is equivalent to the question: who are the authors that have published papers containing sentences that cited references in this dataset?

```
MATCH (a:Author)-[w:PUBLISHES]->
          (p:Paper)-[c:CONTAINS]->
          (s:Sentence)-[d:CITES]-(r:Ref)
RETURN a, w, p, c, s, d
LIMIT 200;
```

Figure 8.2 depicts the result of the Cypher query. The red node is an author. The orange nodes are articles published by the author. The green nodes are sentences in a published article. The purple nodes are references cited by the sentence of the green node connected to them. For instance, the green node at the upper right corner of Fig. 8.2 corresponds to the sentence: "CiteSpace follows a simple model of the dynamics of scholarly communication …" The subject of the sentence is CiteSpace.

More specific questions can be formulated with a Cypher query. For instance, we are interested in authors who published articles that contain statements, or claims, with CiteSpace as the subject. The question is formulated in the Cypher query below.

```
MATCH (c:Claim)<-[r:MAKES]-(s:Sentence)<-[i:CONTAINS]-(p:Paper)<-[j:PUBLISHES]-(a:Author) where c.subject =~ '(?i).*CiteSpace.*' return c,r,s,i,p,j,a;
```

Figure 8.3 shows a visualization of the query result in Neo4j. The visualization shows that the author published two papers that satisfied the criteria. The author is a red node in the graph, connecting to two yellow nodes of papers, which in turn contain sentences in green. The purple nodes are references cited by the green sentences. In other words, the set of green nodes represent citation contexts.

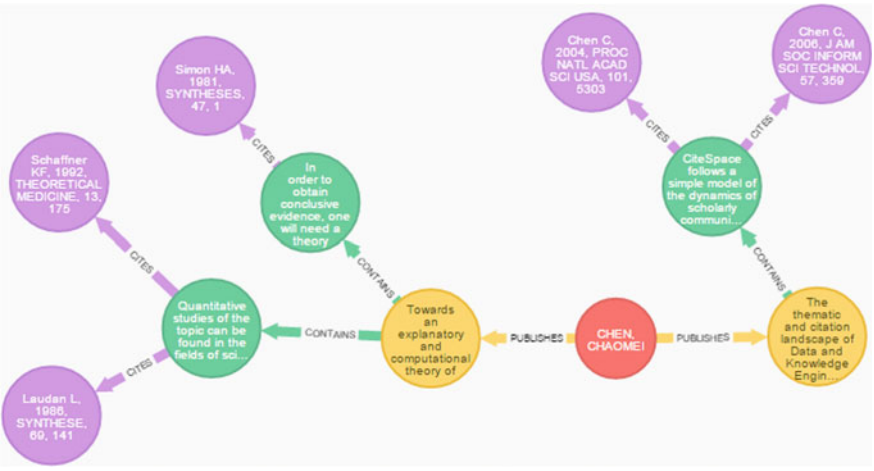

Fig. 8.2 A graphical answer to the question: who has published what paper containing sentences that cited which references?

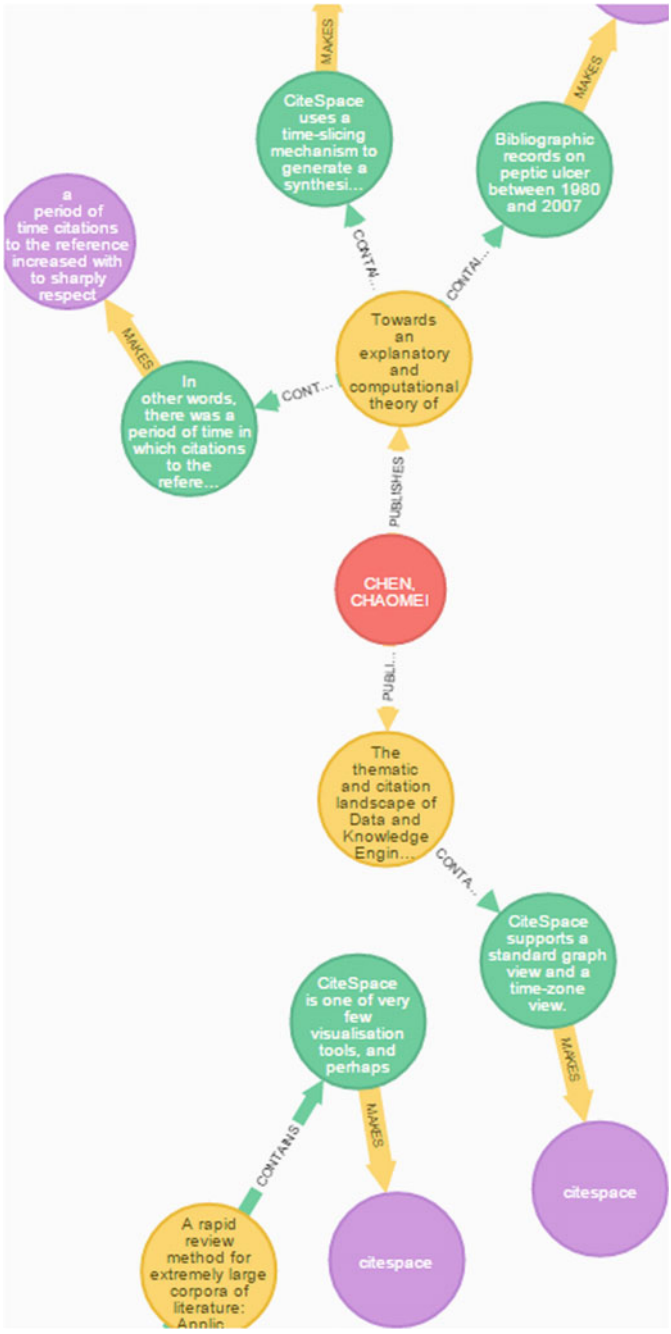

Fig. 8.3 A sub-graph containing sentences in which CiteSpace is the subject

The representation is very flexible. All types of conventional bibliographic networks can be derived from the underlying graph database, including author collaboration networks, citation networks at the author level, article level, and the sentence level.

Table 8.2 includes a list of questions concerning scientific claims made in published articles and corresponding Cypher queries. For example, claims that are associated with hedging words can be identified. Similarly, claims that are associated with uncertain cue words can be also identified.

The following series of MySQL queries are used to collect sentences from MEDLINE articles in the virus dataset and prepare for them to be uploaded to a graph database in Neo4j. We include the completion time on Lenovo W530 so that the reader who is interested can estimate the time required.

Table 8.2 Questions and corresponding queries in Cypher to a graph of scientific publications

Questions in English	Questions in Cypher
When did a claim (hypothesis, assertion, conjecture, statement) appear for the first time in a body of source information, e.g. the literature? In which paper/source?	match (c:Claim)<-[r:MAKES]-(s:Sentence)<-[i:CONTAINS]-(p:Paper)<-[j:PUBLISHES]-(a:Author) where c.object =~ '(?i).*CiteSpace.*' OR c.subject =~ "(?i).*CiteSpace.*' return min(p.year), max(p.year);
How certain was the assertion?	match (c:Claim)-[r]-(s:Sentence)-[:HEDGES]-(w:HedgeTerm) return w.weight;
How many papers made the same claim subsequently?	match (c:Claim)<-[r:MAKES]-(s:Sentence)<-[i:CONTAINS]-(p:Paper)<-[j:PUBLISHES]-(a:Author) where (c.object =~ '(?i).*CiteSpace.*' OR c.subject =~ '(?i).*CiteSpace.*') return count(p);
How many papers made claims that contradict a given claim subsequently?	match (p:Paper)-[:CONTAINS]->(s:Sentence)-[:MAKES]->(d:Claim)-[:CONTRADICTS]-(c:Claim) return count(p);
What are the closely related but distinct claims for a given claim?	match (d:Claim)-[r]-(c:Claim) return d;
Given a claim, which reference is most frequently cited along with the claim's citation context?	match (r:Ref)<-[:CITES]-(s:Sentence)-[:MAKES]->(c:Claim) return r, max(sum(r));
Which references are frequently co-cited in relation to a given claim?	match (s1:Sentence)-[:CITES]->(r1:Ref), (s1:Sentence)-[:CITES]-(r2:Ref), (s1:Sentence)-[:MAKES]->(c:Claim) return r1, r2;
How many dimensions (clusters or eigenvectors or topics) are associated with the citation contexts of a claim?	match (s:Sentence)-[:CITES]->(r:Ref) return s; hierarchical clustering(sentence by word)

```
SELECT *
FROM sentence
INTO OUTFILE 'D:/temp/sentences.csv'
FIELDS ENCLOSED BY '"'
TERMINATED BY ','
LINES TERMINATED BY '\r\n';

Query OK, 143,045,997 rows affected (35 min 39.65 sec)
```

```
CREATE TABLE _virus_sentence
SELECT s.sentence_id AS sid, s.pmid AS pmid,  s.type AS type,
      s.number AS number, s.sentence AS sentence
FROM _virus_year_text AS v, sentence AS s
WHERE v.sid=s.sentence_id;

Query OK, 662132 rows affected (1 hour 59 min 32.65 sec)
Records: 662132  Duplicates: 0  Warnings: 0
```

```
SELECT *
FROM _virus_sentence
INTO OUTFILE 'D:/temp/virus.sentences.csv'
FIELDS ENCLOSED BY '"'
TERMINATED BY ','
LINES TERMINATED BY '\r\n';

Query OK, 662132 rows affected (1.60 sec)
```

The sentences in CSV are loaded to a graph database in Neo4j with the following Cypher queries:

```
CREATE CONSTRAINT ON (s:Sentence) ASSERT s.sid IS UNIQUE;
USING PERIODIC COMMIT 500
LOAD CSV
FROM  'file:///D:/temp/virus.sentences.csv' AS line
WITH LINE LIMIT 662132
MERGE (sentence:Sentence {sid:line[0]})
SET sentence.type=line[2],
    sentence.number=toInt(line[3]),
    sentence.sentence=line[4]
DROP CONSTRAINT ON (s:Sentence) ASSERT s.sid IS UNIQUE;

Added 997459 labels, created 997459 nodes, set 3997459 properties,
statement executed in 449713 ms.
```

The next step is to export the virus dataset of semantic predications from MySQL and then load them to Neo4j. Although it is possible to transfer the data directly from MySQL to Neo4j, it is efficient and reliable to divide the conversion into several smaller steps due to the size of Semantic MEDLINE. The sentence table alone contains 140 million rows.

```
SELECT *
FROM _virus_year_text
INTO OUTFILE 'D:/temp/virus.csv'
FIELDS ENCLOSED BY '"'
TERMINATED BY ','
LINES TERMINATED BY '\r\n';

Query OK, 662132 rows affected (1.60 sec)
```

The semantic predications on virus are uploaded to the Neo4j server with the following Cypher queries. Five types of nodes are added to the graph, namely, paper, sentence, claim, concept, and text. A paper node contains properties such as PMID and the year of publication. A claim corresponds to a semantic predication in Semantic MEDLINE. The subject and object concepts are mapped to concept nodes in the graph. Nodes are connected accordingly based on their types. For example, a paper node CONTAINS a sentence. A sentence MAKES a claim. A subject node as a UMLS concept REPRESENTS its original text.

```
CREATE CONSTRAINT ON (p:Paper) ASSERT p.pmid IS UNIQUE;
CREATE CONSTRAINT ON (s:Sentence) ASSERT s.sid IS UNIQUE;
CREATE CONSTRAINT ON (c:Claim) ASSERT c.pid IS UNIQUE;
CREATE CONSTRAINT ON (a:Concept) ASSERT a.name IS UNIQUE;
CREATE CONSTRAINT ON (t:Text) ASSERT t.text IS UNIQUE;
USING PERIODIC COMMIT 500
LOAD CSV FROM  'file:///D:/temp/virus.csv' AS line
WITH LINE LIMIT 1000000
MERGE (paper:Paper {pmid:line[2], year:toInt(line[8])})
MERGE (sentence:Sentence {sid:line[0]})
MERGE (claim:Claim {pid:line[1], link:line[3]})
      SET claim.subject=line[4], claim.object=line[6]
MERGE (subject:Concept {name:line[4]}) SET subject.type=line[5]
MERGE (object:Concept{name:line[6]}) SET object.type=line[7]
MERGE (s text:Text {text:line[9]})
MERGE (o text:Text {text:line[10]})
MERGE (paper)-[:CONTAINS]->(sentence)
MERGE (sentence)-[:MAKES]->(claim)
MERGE (subject)-[r:CONNECTS {type:line[3]}]->(object)
MERGE (subject)-[:REPRESENTS]->(s text)
MERGE (object)-[:REPRESENTS]->(o text)
DROP CONSTRAINT ON (p:Paper) ASSERT p.pmid IS UNIQUE;
DROP CONSTRAINT ON (s:Sentence) ASSERT s.sid IS UNIQUE;
DROP CONSTRAINT ON (c:Claim) ASSERT q.pid IS UNIQUE;
DROP CONSTRAINT ON (a:Concept) ASSERT a.name IS UNIQUE;
DROP CONSTRAINT ON (t:Text) ASSERT t.text IS UNIQUE;
```

The process took 748,162 ms to complete. The resultant graph contains 1.07 million nodes, 4.29 million properties, and 1.39 million relationships. It takes up 1.8 GB of disk space on the computer. As shown in Table 8.3, the semantic graph on virus features 553,169 sentences from 320,818 MEDLINE articles. These sentences collectively make 136,209 claims, i.e., semantic predications, involving 18,723 UMLS concepts, which in turn represent 66,584 words or phrases in the original unstructured text.

We will illustrate the usage of the virus graph with a few examples. One is to find claims of causality concerning Ebola.

Table 8.3 The graph constructed from the semantic predications on virus

Node type	Node count	Link type	Link count
Paper	320,818	CONTAINS	1,106,338
Sentence	553,169	MAKES	1,255,560
Claim	136,209		
Concept	18,723	CONNECTS	271,703
Text	66,584	REPRESENTS	140,858

Causality Claims on Ebola

Claims that identify how one entity may influence another are defined as causality claims, for example, as in the claim that Ebola virus ∼ CAUSES ∼ Hermorrhagic fever. The response time of the following Cypher query is less than 0.5 s. The Cypher query is asking for claim nodes such that the Ebola concept is either the subject or the object. The resultant claims are shown in Table 8.4, where c.pid is the semantic predication ID in Semantic MEDLINE.

```
match (c:Claim) where c.subject =~ '(?i).*ebola.*' or c.object =~
'(?i).*ebola.*' return c.pid, c.subject, c.link, c.object order by c.link limit
10;

Returned 10 rows in 477 ms.
```

If we want to dig deeper, we can ask which papers made these claims with the following Cypher query. This query takes slightly over 1 s to complete (1026 ms to be precise).

```
match (p:Paper)-[r*1..3]->(c:Claim)
where c.subject =~ '(?i).*ebola.*' or c.object =~ '(?i).*ebola.*'
return p.pmid, p.year, c.pid, c.subject, c.link, c.object order by c.link limit 10;
```

Table 8.4 Causality claims concerning Ebola

c.pid	c.subject	c.link	c.object
14126539	Mutation	AFFECTS	Ebola virus
1705776	Ebola virus	CAUSES	Hemorrhagic Fevers, Viral
6248254	Ebola virus	CAUSES	Disease
7722924	Ebola virus	CAUSES	Hemorrhagic Fever, Ebola
7481712	Ebola virus	CAUSES	Hemorrhagic Disorders
7351555	Ebola virus	CAUSES	Acute Disease
9991484	Ebola Virus, Sudan	CAUSES	Communicable Diseases
9991327	Ebola Virus, Zaire	CAUSES	Disease
9501704	Ebola virus	CAUSES	Infection
9050015	Ebola virus	CAUSES	Laboratory Infection

A question that is probably more relevant to the purpose of identifying emerging trends would be: who and which paper was the first to make claims on Ebola? As shown in Table 8.5, the earliest causality claim involving Ebola first appeared in 1978. The claim that Ebola virus causes viral hemorrhagic fevers appeared first in a 1978 MEDLINE article (PMID 352653).

```
match (p:Paper)-[r*1..3]->(c:Claim)
where c.subject =~ '(?i).*ebola.*' or c.object =~ '(?i).*ebola.*'
return p.pmid, p.year, c.pid, c.subject, c.link, c.object order by p.year limit 5;
```

The multi-type relationships among articles, claims, and semantic types are visualized in Fig. 8.4 through a built-in visualization function in Neo4j. To generate the visualization, the user just needs to formulate a Cypher query to specify the conditions to be satisfied.

Table 8.5 Papers that made the earliest causality claims on Ebola

p.pmid	p. year	c.pid	c.subject	c.link	c.object
307455	1978	1650932	Immunofluorescence	DIAGNOSES	Ebola virus
352653	1978	1705776	Ebola virus	CAUSES	Hemorrhagic Fevers, Viral
503930	1979	1705776	Ebola virus	CAUSES	Hemorrhagic Fevers, Viral
119829	1979	2294309	Vero Cells	LOCATION_OF	Ebola virus
94087	1979	2294309	Vero Cells	LOCATION_OF	Ebola virus

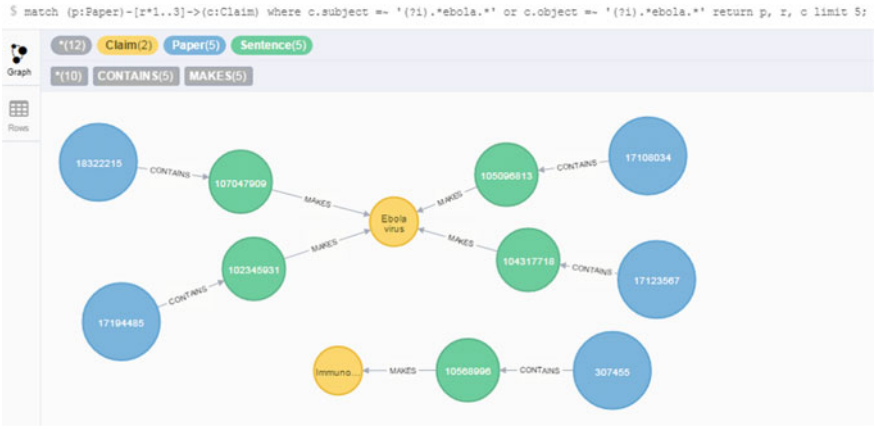

Fig. 8.4 The earliest causality claims involving Ebola

Conflicting Claims

Conflicting claims are a major source of uncertainty in scientific literature. The hallmark of a domain expert is the ability to differentiate conflicting information and contradictory claims. In addition to the positive causality claims shown in Table 8.5, there are claims that negate the causality associated with Ebola (Table 8.6). The claim that virus NEG_CAUSES Ebola Hemorrhagic Fever is extracted from the sentence: "These results suggest that the different clinical outcomes of EBOV infection do not result from virus mutations." By contrasting contradictory claims, one may identify provenance of evidence associated with such contradictions and track the developments that may lead to a reconciliation. A valuable practical potential of this capability is to monitor claims made by retracted papers and to detect the consistency between a given paper and the rest of the body of knowledge.

When Was a Causal Relationship Initially Hypothesized?

Researchers often need to trace to the origin of a hypothesis in the literature. The following example illustrates how we can query when and which paper hypothesized a causal relationship based on the semantic graph. The Cypher query below retrieves chains of papers ∼ sentences ∼ claims such that the sentences contain the word 'hypothesi,' and the semantic links are one of the casual relations. The partial word 'hypothesi' is used to catch variations such as hypothesis and hypothesized,

```
match (p:Paper)-[]-(s:Sentence)-[]-(c:Claim)
where s.sentence<>'' and s.sentence =~ '(?i).*hypothesi.*' and c.link =~
    '(?i).*CAUSES|AFFECTS|TREATS|INHIBITS|DISRUPTS|PREVENTS|PRED
    ISPOSES|CONVERTS_TO'
return p.pmid, p.year, s.type,s.sentence, c.subject,c.link, c.object
LIMIT 15;

Returned 9 rows in 2698 ms.
```

Table 8.6 Negations of causality claims on Ebola

p.pmid	p. year	c.subject	c.link	c.object
11752702	2002	Virus	NEG_CAUSES	Hemorrhagic Fever, Ebola
17940961	2007	Ebola Virus, Zaire	NEG_CAUSES	Disease

Table 8.7 illustrates two examples from titles, whereas Table 8.8 shows two examples from abstracts. For example, the idea that a virus can cause cancer first appeared in a 1967 article (PMID 5596354).

Now we know the existence of two claims, or semantic predications, that virus CAUSES Malignant Neoplasms, and that virus CAUSES Neoplasm. Were these predications discussed in other MEDLINE articles? Given a semantic predication, show us the articles associated with the predication. Suppose we are interested in the latter predication, which has a predication ID of 544471 in Semantic MEDLINE. The Cypher query below retrieves all the articles with sentences connected to the predication.

```
match (p:Paper)-[]-(s:Sentence)-[]-(c:Claim{pid:'544471'})
where s.sentence =~ '(?i).*' and
        c.link =~ '(?i)CAUSES|INTERACTS_WITH|AFFECTS|PREVENTS|TREATS'
return count(distinct(p)), count(distinct(s)), count(distinct(c));
```

Table 8.9 shows the results of the Cypher query. A total of 118 articles are connected to the predication 544,471. The earliest appearance was 1926.

Measuring the Importance of Semantic Predications

One way to identify important semantic predications is based on whether they have attracted a lot more attention than their peers from the scientific community. From a sociological perspective of scientific change (Fuchs 1993), researchers are driven by their competitions for recognition and reputation. A topic that attracts much attention from researchers is apparently considered important. The uncertainty associated with a high-attention topic must be also high—there must be a lot of potential to make high-reward discoveries or something that can dramatically boost one's recognition or reputation. Thus, competitions in high-profile and high-risk areas of research tend to be more intensive than other areas. In contrast, research areas with a low level of uncertainty are unlikely to sustain intensive competitions.

Table 8.7 MEDLINE articles that hypothesized causal relations in titles

PMID	Year	Sentence	Subject	Relation	Object
5596354	1967	[Hypothesis that cancer can be caused by a virus].	Virus	CAUSES	Malignant Neoplasms
1435387	1992	Human cancers and viruses: a hypothesis for immune destruction of tumours caused by certain enveloped viruses using modified viral antigens.	Virus	CAUSES	Neoplasm

Table 8.8 MEDLINE articles that hypothesized causal relations in abstracts

PMID	Year	Sentence	Subject	Relation	Object
8780661	1996	It was hypothesized that if the transmission of CMV through transfusion causes CMV disease in human immunodeficiency virus-positive hemophiliacs, then hemophiliacs with CMV AIDS would be more likely to have received transfusions than those with AIDS-defining disease not caused by CMV (non-CMV AIDS)	Cytomegalovirus	CAUSES	Disease
17429926	2007	The objective of this study was to investigate the association hypothesis that outcome following respiratory syncytial virus (RSV) induced bronchiolitis (RSVB) and RSV induced wheeze (RSVW) are different	Respiratory syncytial virus	CAUSES	Bronchiolitis

The reasoning concerning attention, uncertainty, risk, potential reward, and ultimate recognition suggests that we may learn valuable insights from semantic predications that stand out in how long and how much of attention they have ever generated in scientific articles published in the past. Burst detection (Kleinberg 2002) is a generic algorithm that can be used to identify the level of attention as a type of burstness over time. If a semantic predication appears at a much higher level of frequency than other semantic predications within the same research domain, then intuitively, the semantic predication is having a burst. The duration in which a relatively high-level frequency is observed defines a duration of burstness. If multiple levels of frequency are observed in association with a semantic predication, then the predication may experience hierarchically related bursts. For example, a semantic predication concerning the semantic relation between virus and infection

Table 8.9 Earliest sentences concerning the predication: Virus CAUSES Neoplasm (PID: 544471)

PMID	Year	SID	Sentence
19869151	1926	6276721	Approximately 1 c.mm. of spleen tissue in 3,000 c.mm. of medium may on occasion maintain a concentration of Rous virus in this fluid sufficient to produce a tumor upon inoculation into chickens
19870262	1934	5867204	It is often impossible to determine whether the neoplasms caused by the virus of Strain 2 are of endothelial or mesenchymal origin, and it is possible that both types of cells may be stimulated by the same virus
19870455	1936	5451445	This virus produces neoplasms only when brought in contact with bone or cartilage
21001044	1946	5828267	Title: Induction of neoplasia in vitro with a virus
13185069	1953	6779315	Title: [Cellular multiplication and tumors induced by virus; cancer as an infection]

may have a level of frequency higher than some relatively rare diseases in the Semantic MEDLINE, thus, the predication may have a long period of relatively low-level burstness. From time to time, the semantic predication may have an even higher level of frequency, e.g., when particular types of virus, such as HIV, H1N1, and SARS are involved.

CiteSpace provides a simple user interface for applying the burst detection to semantic predications (Fig. 8.5). The user can specify the semantic types of semantic predications. In the following example, causal relations such as CAUSES, AFFECTS, INHIBITES, DISRUPTS, PRODUCES, and PREVENTS are specified. Parameters for the burst detection model are set so that a burst must last at least for three consecutive years between 1914 and 2014, a 101-year time span.

Kleinberg's algorithm found 167 qualified semantic predications. Figure 8.6 illustrates some of the earliest ones. These predications are considered particularly important by scientists who publish in biomedicine. The darker blue bars depict the period in which a semantic predication appears in MEDLINE for the first time. Each red bar depicts the duration of a burst, which is a period in which the frequency of the occurrences of the predication is considerably higher relative to that of other predications at the same time. The overall profile of the red lines suggests that (1) the burstness of a predication tends to shift over time and (2) the burstness of a predication becomes relatively short except the first few long-lasting ones. Thus, the immediate conclusion is that researchers' focus changes over time. This observation can be used as the basis of detecting emerging topics at a large scale.

Table 8.10 shows a more selective group of semantic predications that have the strongest bursts (burst strength > 10.0) among the 167 predications with bursts. The earliest predication, Virus CAUSES Influenza, first appeared in 1918, which would immediately remind us the 1918–1919 Spanish Flu. A particularly strong burst, highlighted in Table 8.2, is found with the *HIV => Acquired Immunodeficiency*

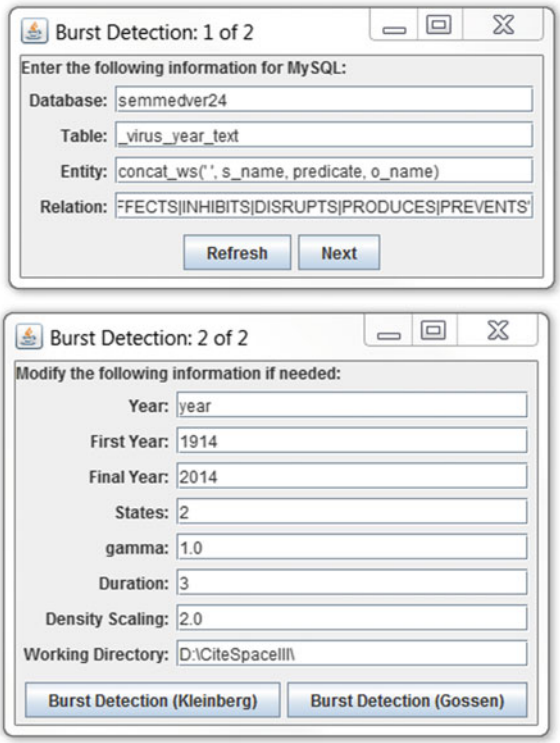

Fig. 8.5 Detecting bursts in semantic predications on causal relations in research on virus

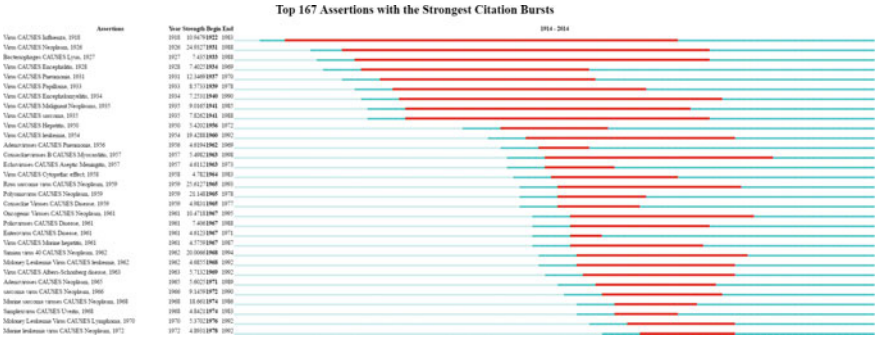

Fig. 8.6 Some of the earliest semantic predications found with bursts

Syndrome predication, which first appeared in 1984 and started to burst for 10 years from 1991 to 2000. The strongest burst belongs to the predication *Human Papillomavirus => Malignant neoplasm of cervix uteri*, which first appeared in 1986. It didn't burst until 26 years later in 2011. It is worth noting that the Nobel

Table 8.10 Semantic predications with burstness strengths > 10.0 from top 167 ones with bursts. '=>' denotes CAUSES

Assertions	Strength	Begin	End	1914 - 2014
Virus => Influenza, 1918	10.9479	**1922**	1983	
Virus => Neoplasm, 1926	24.9327	**1931**	1988	
Virus => Pneumonia, 1931	12.3469	**1937**	1970	
Virus => leukemia, 1954	19.4288	**1960**	1992	
Rous sarcoma virus => Neoplasm, 1959	25.6127	**1965**	1993	
Polyomavirus => Neoplasm, 1959	21.148	**1965**	1978	
Oncogenic Viruses => Neoplasm, 1961	10.4718	**1967**	1995	
Simian virus 40 => Neoplasm, 1962	20.0066	**1968**	1994	
Murine sarcoma viruses => Neoplasm, 1968	18.661	**1974**	1986	
Avian Sarcoma Viruses => Neoplasm, 1975	10.6576	**1981**	1991	
Virus => Disease, 1922	11.5348	**1981**	1984	
Virus => Acute Erythroblastic Leukemia, 1977	10.9621	**1983**	1998	
Cytomegalovirus => Pneumonia, 1975	11.1421	**1985**	1997	
Virus => Acquired Immunodeficiency Syndrome, 1983	16.5985	**1989**	1995	
Retroviridae => Acquired Immunodeficiency Syndrome, 1984	23.2685	**1990**	1996	
HIV => Acquired Immunodeficiency Syndrome, 1984	38.7063	**1991**	2000	
Retroviridae => Disease, 1982	11.4849	**1991**	1998	
HIV => Infection, 1985	24.7191	**1993**	2001	
Hepatitis C virus => Hepatitis C, 1989	13.3138	**1996**	2005	
HIV-1 => Infection, 1989	12.5993	**1998**	2004	
Kaposi Sarcoma AFFECTS human herpesvirus 8, 1996	17.1303	**2003**	2010	
Hepatitis C virus => Chronic liver disease NOS, 1989	14.1651	**2004**	2009	
human herpesvirus 8 => Kaposi Sarcoma, 1996	24.5536	**2004**	2009	
Respiratory syncytial virus => Lower respiratory tract infection, 1986	14.6341	**2005**	2014	
Hepatitis C virus => Fibrosis, 1989	14.974	**2007**	2014	
Hepatitis C virus => Primary carcinoma of the liver cells, 1990	12.7617	**2007**	2014	
West Nile virus => Encephalitis, 1969	11.8932	**2009**	2014	
SARS coronavirus => Severe Acute Respiratory Syndrome, 2003	17.8423	**2009**	2014	
Norovirus => Gastroenteritis, 1999	25.8451	**2010**	2014	
Human Metapneumovirus => Respiratory Tract Infections, 2003	12.9203	**2010**	2014	
Norovirus => Acute gastroenteritis, 2000	19.819	**2010**	2014	
Human Papillomavirus=>Malignant neoplasm of cervix uteri, 1986	42.2078	**2011**	2014	
Human Papillomavirus => Anogenital venereal warts, 1983	16.2154	**2012**	2014	

Prize in Physiology or Medicine in 2008[1] was awarded to Harald zur Hausen for his discovery of "human papilloma viruses causing cervical cancer" and the other half of the prize was awarded jointly to Françoise Barré-Sinoussi and Luc Montagnier for their discovery of "human immunodeficiency virus"—the HIV. The two predications with the strongest burstness are both Nobel Prize winning topics!

Contradictions as a Source of Uncertainty

Research fronts are typically involved with a high-level of uncertainty, where research questions have yet answered and controversial findings have yet settled. Scientific claims surrounded by cues that indicate the involvement of incomplete, conflicting, and contradictory information point to areas of research where the uncertainty is high and the competition is likely worthwhile.

Funding agencies' decisions on high-risk and high-reward research are often revealing in terms of how people make decisions involving a significant degree of

[1] https://www.nobelprize.org/nobel_prizes/medicine/laureates/2008/press.html.

uncertainty. Wagner and Alexander (2013) evaluated the Small Grants for Exploratory Research (SGER) program of the U.S. National Science Foundation (NSF). SGER was a 16-year program operated from 1990 until 2006. The program was designed to serve as a special funding device to support high-risk and high-reward research that is unlikely to get funded through the traditional evaluation system. Citation counts, expert interviews, and the results of a survey all indicate a successful SGER. On the other hand, the evaluation of the program reveals that the NSF program directors were perhaps overly conservative—they spent far less than the allowable funds that were ear-marked for exploratory research. The program directors remained risk averse even with a program particularly designed to encourage transformative research. Similarly, Laudel and Glaser (2014) studied links between epistemic properties and institutional conditions for research based on projects funded by the European Research Council (ERC). They found that research that is important for the progress of a field is difficult to fund with common project grants. The conventional funding mechanisms appear to discourage unconventional research across all disciplines.

The perceived risk is in part due to the epistemic uncertainty—the scientific community is simply lacking the knowledge to remove the uncertainty or the controversies. Semantic predications and their original text in Semantic MEDLINE provide a useful resource for the study of scientific claims along with the extent to which conflicting or contradictory information is involved.

Each semantic predication consists of a subject, a predicate, and an object. If we can identify those predicates that are particularly associated with sentences in scientific articles that contain indicators of conflicting or contradictory information, then we may reach a better understanding of their patterns.

We constructed two sets of sentences: one set contains sentences that include terms such as conflicting, contradictory, and inconsistent; the other set contains sentences that do not include such terms. We call the former set the conflict set and the latter the no conflict set. Next, we compare the semantic types of the semantic predications associated with each of the two sets.

Figure 8.7 is a log-log plot of predicates found in the no conflict set (x-axes) against their frequencies in the conflict set. The size of a node represents the frequency of the corresponding predicate. Predicates in dark red are those involved in causal relations, such as TREATS, AFFECTS, and CAUSES, whereas those in light red are involved in structural or ontological relations, such as PROCESS OF, IS A, and PART OF. The line divides predicates into two parts. Predicates above the line, such as, TREATS, AFFECTS, and PREVENTS, appeared more often in the conflict set, whereas predicates below the line appeared more often in the no-conflict set, such as CAUSES, INHIBITS, and DISRUPTS. In both groups, predicates of causal relations are dominating the overall semantic predications. The conflict set appears to have more active or positive predications, i.e., the A causes B pattern. In contrast, the no-conflict set appears to have more passive or negative predications, i.e., the A suppresses B pattern.

Figure 8.7 also shows that the conflict set contains relatively more semantic predications with the semantic types such as PROCESS_OF, IS_A,

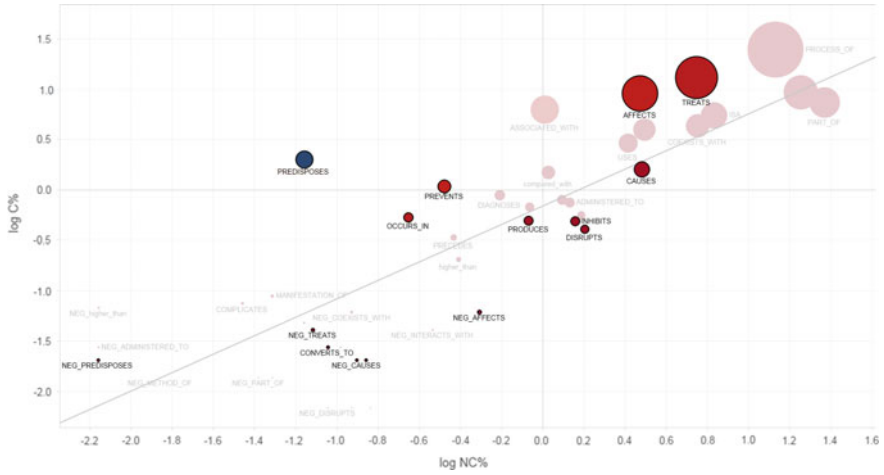

Fig. 8.7 Major semantic relations in the conflict versus no conflict sets of articles from Semantic Medline

ASSOCIATED_WITH, and USES as well as TREATS, AFFECTSS, and other types of relations. The division suggests the richness of the conflict set and the overall significance of the research topics that are intrinsically more essential to research.

Semantic Predications on Virus Research (1914–2014)

We constructed a subset of semantic predications on virus from Semantic MEDLINE (version: semmedver24) so that we can explore semantic predications closely in terms of their various properties associated with representing scientific knowledge, including advantages and weaknesses.

The virus subset contains semantic predications and their original text as long as either the subject or the object of a predication has the semantic type of virus, which is the name used in MEDLINE.

The MySQL query to extract the subset and the status report are as follows:

```
CREATE TABLE _virus
SELECT count(*)
FROM predication_aggregate
WHERE s_type = 'virs' OR o_type='virs';

Query OK, 638792 rows affected (54.92 sec)
```

A total of 638,792 qualified predications were selected from the table predication_aggregate. The citation table contains 23.7 million MEDLINE articles

(23,657,386 for sememdver24). Linking the two large tables with a MySQL Join took 52 min on a Lenovo Workstation W530. Among the 638,792 predications, about 21.3% (136,214) are unique predications, which involve 28 types of relations.

To reduce the need for possibly time-consuming table joins in the future, we created a temporary table that contains all the major information about predications along with the year of publication for the original article. Adding the time, i.e. the year of publication, allows us to perform algorithms such as burst detection, which identifies which semantic predications are attracting the attention of researchers in terms of how fast their frequencies increase in the literature. The following query joined two tables so that each predication can be timestamped by the year of the original article's publication. The new table contained 638,780 predications.

```
CREATE TABLE _virus_year
SELECT pid, sid, a.pmid as pmid,
     predicate, s_name, s_type, o_name,
     o_type, b.pyear as year
FROM
    _virus AS a,
    citations AS b
WHERE a.pmid=b.pmid;

Query OK, 638780 rows affected (52 min 33.63 sec)
Records: 638780  Duplicates: 0  Warnings: 0
```

Figure 8.8 depicts the distribution of the number of semantic predications per year. Since 1975, there are 5000 predications each year. The 5-year moving average is closely tracking the top of the bars. In 2013, the number of the predications is over 30,000.

The distribution of the semantic predications on virus research shows that the majority of the predications appeared after 1975 and there were more predications in early 1990s than the value of a 5-year moving average.

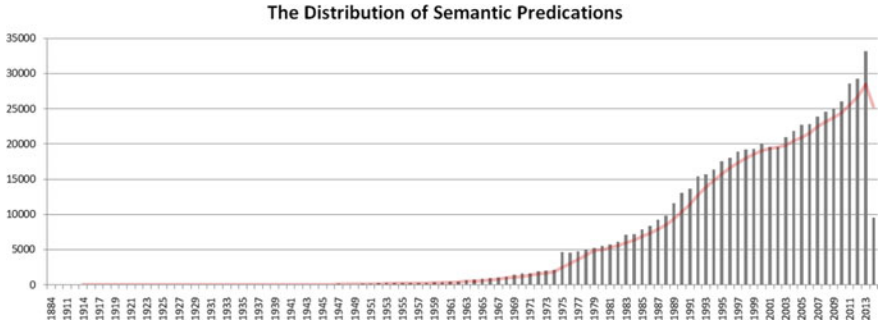

Fig. 8.8 The distribution of semantic predications over time. The red line is the 5-year moving average

The third query aims to match specific information from a sentence with a given predication. On a Lenovo W530 workstation, it took 2 h to complete the query.

```
CREATE TABLE _virus_year_text
SELECT a.sid as sid, a.pid as pid, a.pmid as pmid, a.predicate as
        predicate, a.s_name, a.s_type, a.o_name, a.o_type, a.year,
        b.subject_text, b.object_text
FROM _virus_year AS a, sentence_predication AS b
WHERE a.sid=b.sentence_id AND a.pid=b.predication_id;

Query OK, 662132 rows affected (2 hours 2 min 8.92 sec)
Records: 662132  Duplicates: 0  Warnings: 0
```

There are a total of 136,209 distinct semantic predications in the _virus_year_text table. The following query can be used to find predications that appeared most frequently in the subset of virus:

```
SELECT
    b.c, -log2(b.c/136209),
    b.pid, b.predicate,
   concat(b.s_name, ' ', b.predicate, ' ', b.o_name)
FROM (
          SELECT count(*) AS c, pid,
             s_name,
             predicate,
             o_name
          FROM _virus_year_text
  GROUP BY pid
  ORDER BY count(*) DESC
) AS b
LIMIT 10;
```

Table 8.11 lists the top 10 most frequently appeared semantic predications in the virus subset. The top 10 predications' predicates include three PROCESS OF, six PART OF, and one LOCATION OF.

The information content of most frequently appeared semantic predications are lower than low-frequency predications. Predications that are relative rare have high information contents. Table 8.12 is generated with the following query, which searches for predications that have a particular number of appearances, e.g. b.c = 10 for 10 appearances, in the virus dataset. The query also calculates the IC of each predication based on the total of 136,209 predications. In addition, predicates are limited to predicates with implications of causal relations such as AFFECTS, CAUSES, TREATS, INHIBITES, and DISRUPTS.

Table 8.11 Top 10 most frequently appeared semantic predications in the virus subset

Count	IC	PID	Predicate	Predication
26077	2.3850	2634059	PROCESS_OF	Communicable Diseases PROCESS OF Hepatitis C virus
6076	4.4866	1666937	PART_OF	DNA PART OF Human Papillomavirus
5456	4.6418	946718	PROCESS_OF	Disease PROCESS OF Cytomegalovirus
5111	4.7361	2383142	PROCESS_OF	Communicable Diseases PROCESS OF HIV-1
4474	4.9281	2407960	PART_OF	RNA PART OF HIV
3764	5.1774	541931	LOCATION_OF	Cells LOCATION OF Virus
3331	5.3537	3027212	PART_OF	Vaccines PART OF Human Papillomavirus
3318	5.3594	4640493	PART_OF	RNA PART OF Hepatitis C Virus
2578	5.7234	1467311	PART_OF	Large T-Antigen PART OF Simian virus 40
2221	5.9385	663163	PART_OF	DNA PART OF Simian virus 40

Table 8.12 Some examples of rare predications with 1, 5, or 10 appearances in total

Count	IC	PID	Predicate	Predication
10	13.7335	608940	CAUSES	Enterovirus CAUSES Syndrome
10	13.7335	764171	AFFECTS	Infection AFFECTS Virus
10	13.7335	1076906	CAUSES	Virus CAUSES Latent Infection
10	13.7335	1138218	CAUSES	Papillomavirus, Cottontail Rabbit CAUSES Papilloma bit CAUSES
10	13.7335	1238699	CAUSES	Enterovirus CAUSES Conjunctivitis, Acute Hemorrhagic
5	14.7335	588468	CAUSES	Adenoviruses CAUSES Pharyngo-Conjunctival Fever
5	14.7335	652125	CAUSES	ECHOVIRUS 11 CAUSES Disease
5	14.7335	695371	CAUSES	Simplexvirus CAUSES Primary infection NOS
5	14.7335	740895	CAUSES	Fibroma Virus, Rabbit CAUSES Neoplasm
5	14.7335	771239	CAUSES	Virus CAUSES Tick-Borne Encephalitis
1	17.0556	541848	AFFECTS	Carcinoma AFFECTS Tick-Borne Encephalitis Virus
1	17.0556	543509	CAUSES	Rabies virus CAUSES Multiple Sclerosis
1	17.0556	547396	CAUSES	Echoviruses CAUSES Meningococcal meningitis
1	17.0556	550507	CAUSES	Mumps virus CAUSES comatose
1	17.0556	558749	CAUSES	sarcoma virus CAUSES Malignant Neoplasms

```
SELECT
      b.c,
      -log2(b.c/136209),
      b.pid,
      b.predicate,
      concat(b.s_name, ' ', b.predicate, ' ', b.o_name)
FROM (
         SELECT
            count(*) AS c,
            pid, s_name, predicate, o_name
         FROM _virus_year_text
         WHERE
              predicate regexp
                'AFFECTS|CAUSES|TREATS|INHIBES|DISRUPTS'
         GROUP BY pid
         ORDER BY count(*) DESC
      ) AS b
WHERE b.c=10
LIMIT 5;
```

We can export the content of the table to a comma separated values (CSV) file with the following query:

```
SELECT count(*), pid, predicate
FROM _virus_year_text
GROUP BY pid
INTO OUTFILE 'D:/temp/frequencies_of_pids.csv'
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\r\n';
```

Exploring a Semantic Network of Predications in CiteSpace

A collection of semantic predications forms a network with UMLS concepts as nodes and semantic types as relations. Given the variety of visual analytic functions provided by CiteSpace, structural and temporal patterns in a set of semantic predications can be studied as an associative network. For example, a semantic network of UMLS concepts and their semantic relations can be constructed from a given set of semantic predications. Similarly, as shown at the beginning of the chapter, one can also construct a graph database and explore various graph-theoretical questions in graph database query languages such as Cypher in Neo4J.

Causal Relations in Virus Research

As summarized in Table 8.13, the total of 662,132 instances of semantic predications concerning a virus in one way or another came from 320,818 MEDLINE articles. The number of unique predications is 136,209. On average, each

Table 8.13 Statistics of semantic predications concerning viruses

	Semantic predications		MEDLINE articles
	Total	Unique	Unique
Virus	662,132	136,209	320,818
Causal relations in virus	50,861	15,902	38,256

predication is expected to appear five times, although we know its distribution is skewed. There are 15,902 unique semantic predications are related to assertions on causal relations, such as HIV CAUSES AIDS. There are a total of 50,861 instances of these causal predications from 38,256 MEDLINE articles.

Table 8.14 shows top 20 most popular types of predicates in the set of predications on virus. Predicates such as PART_OF, PROCESS_OF, LOCATION_OF, and IS_A are essential to ontological structures, whereas predicates such as CAUSES, INTERACTS_WITH, AFFECTS, and PREVENTS are assertions concerning the impact of one concept on another or changes that one may cause in the other. Predicates on the second half of the table are a series of predication types that negate those in the first half. NEG_CAUSES, for example, negates the predicate CAUSES as in HIV NEG_CAUSES AIDS, which would be equivalent to the assertion that HIV does not cause AIDS.

Table 8.14 Top 20 most popular types of predicates in the virus dataset

Count	Predicate
248756	PART_OF
163969	PROCESS_OF
111078	LOCATION_OF
41860	CAUSES
27942	ISA
24062	INTERACTS_WITH
19361	DIAGNOSES
12807	COEXISTS_WITH
6283	AFFECTS
1838	PREVENTS
910	NEG_LOCATION_OF
783	NEG_INTERACTS_WITH
632	NEG_PART_OF
611	NEG_CAUSES
386	NEG_PROCESS_OF
293	NEG_COEXISTS_WITH
239	NEG_DIAGNOSES
204	NEG_AFFECTS
53	NEG_PREVENTS
35	compared_with

In CiteSpace, under the Data menu, there is an item Semantic MEDLINE > Semmed2WoS. This function executes the following query to retrieve predications in which the subject causes changes in the object. In particular, several relations meet this condition, namely CAUSES, TREATS, AFFECTS, PREVENTS, INHIBITS, and INTERRUPTS. Each of the relations specifies a change induced by the subject of the predication. In addition, we are also interested in the negation of such relations, for example, NEG_CAUSES and NEG_AFFECTS because of the importance of knowledge concerning the causality.

```
SELECT *
FROM _virus_year_text
WHERE predicate REGEXP
   'CAUSES|TREATS|AFFECTS|PREVENTS|
    NEG_CAUSES|NEG_AFFECTS|NEG_PREVENTS|
    INHIBITS|INTERRUPTS';
```

The above query found 38,256 MEDLINE articles containing 50,861 semantic predications on causal relations, which represent 15,902 unique semantic predications. CiteSpace converts these MEDLINE records to a format similar to the Web of Science such that the user can use CiteSpace's visual analytic functions to explore the structure and dynamics of these predications over time (Chen 2017). The user can simplify the network with functions such as Pathfinder network scaling and analyze transformative potentials of MEDLINE articles through Structural Variation Analysis (Chen 2012).

Semantic predications of a MEDLINE article are converted to a format that extends the standard Web of Science format (Fig. 8.9). For example, an article (PMID: 24099575) published in 2013 contains four semantic predications. These predications are mapped to an extended field XX, which can be recognized by CiteSpace to visualize such records as part of a network of concepts linked by corresponding semantic predications. The number of predications is set as the value of the TC field, which can be used as basis for selecting articles based on how many distinct predications they have.

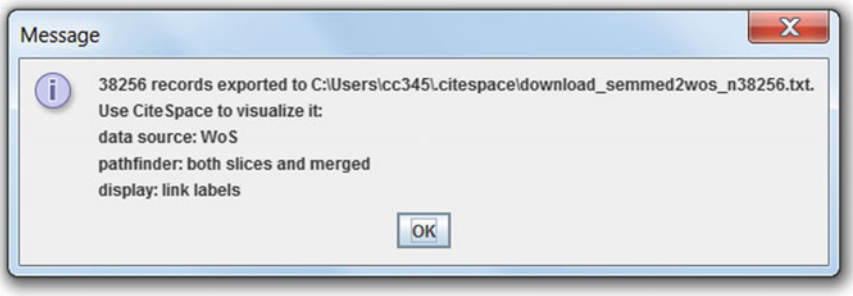

Fig. 8.9 A total of 38,256 MEDLINE records are converted to a data file for subsequent analysis with CiteSpace

```
PT J
TI Rabbit Hemorrhagic Disease Virus CAUSES RHD
SO Medline PMID 24099575
DT Article
DE Rabbit_Hemorrhagic_Disease_Virus~CAUSES~Rheumatic_Heart_Disease;
   European_brown_hare_syndrome_virus~CAUSES~Hepatitis;
   Rabbit_Hemorrhagic_Disease_Virus~CAUSES~Hepatitis; Lagovirus~CAUSES~Hepatitis
NR 0
TC 4
PD JUN-15
PY 2013
PM 24099575
XX rabbit hemorrhagic disease virus      rhd       0.03886223059344139    CAUSES
  european brown hare syndrome virus     hepatitis 0.03886223059344139    CAUSES
  rabbit hemorrhagic disease virus       hepatitis 0.03886223059344139    CAUSES
  lagovirus    hepatitis 0.03886223059344139    CAUSES
ER
```

Visual Analysis of Semantic Predications

In a network of co-cited references, each node is a scientific publication. Two nodes are connected if a subsequently published article cited both of them within the article. The network can be divided into clusters, or groups of references, such that each group can be characterized by some themes. Furthermore, these themes are typically shared by references within the same cluster, but it is less likely to be shared with references in other clusters (Chen et al. 2010; Chen 2017). Each cluster is resultant from the work of a specialty, i.e. a community of researchers who tend to publish in a set of interrelated journals or bump into each other at conferences they regularly attend. We can identify an important article by its citation burst and we will attribute the most significant contributions made by the article to the article as a whole. Thus each article serves a role that is much like a concept. Small (1978), who pioneered much of the co-citation analysis methodology, coined the term concept symbols.

The resolution of a GPS device determines the extent to which it can locate a position with confidence. It becomes helpless if the precision required to accomplish a task is below the finest level of granularity the GPS can reach. Scientometrics at the granularity of an article level can answer many of our questions. However, with the resolution at the article level, it is difficult for us to address many more specific questions. A semantic predication represents a semantic relation between two concepts. Usually, one of the concepts is called the subject and the other is called the object. The semantic relation represents how the subject and the object are connected semantically. For example, “HIV causes AIDS” is semantic predication. HIV is the subject, whereas AIDS is the object. The verb causes is the semantic relation.

There are distinct advantages of representing the knowledge of a scientific domain in terms of semantic predications. Semantic predications provide more precise representations of knowledge than using articles as a whole.

To what extent is the methodology that we have demonstrated at the article level applicable to the study of a scientific domain at the level of semantic predications? We will adapt the methodology and apply it to the study of virus research in the following example.

The source of input data is MEDLINE. Similar to a bibliographic record in sources such as the Web of Science, a MEDLINE record includes the meta-data of a scientific publication, including the title, the abstract, and a list of keywords. Unlike a record in the Web of Science, a MEDLINE record is indexed by a number of MeSH terms—Medical Subject Headings. MeSH terms are from the controlled vocabulary thesaurus compiled by the U.S. National Library of Medicine. MeSH terms are organized in a hierarchy. Unlike the Web of Science records, MEDLINE records do not include information on cited references. There are two ways to obtain semantic predications from scientific publications, primarily within the scope of biomedicine research. One is to extract semantic predications by using SemRep and the other is to use semantic predications extracted by the SemMed project. We have introduced both SemRep and SemMed in Chap. 5.

Each MEDLINE record may have one or more semantic predications. Semantic predications from the same MEDLINE record are co-occurring predications. As we have seen, a set of co-occurring entities can be represented as a network of interconnecting entities. We assign a timestamp to each predication. The timestamp registers the time when the semantic predication appears in our dataset for the first time. Thus we can treat the collection of semantic predications in the same way as we treat co-occurring keywords in the Web of Science records. For example, we can run a burst detection to see which semantic predications have abrupt increases in their occurrences. We can divide a network of semantic predications into clusters so that we can see which semantic predications tend to be discussed together. We can generate timeline visualizations and see how they evolve over time. We can perform a structural variation analysis and identify novel connections between semantic predications. In other words, we can apply many analytic techniques developed for document co-citation analysis to semantic predications.

Constructing a Semantic Network

Unlike bibliography records in the Web of Science or Scopus, a MEDLINE article does not include references cited by the article. When we construct a network of cited references in CiteSpace, a common strategy is to select articles that have been cited to an extent themselves and build the network of references cited by these elite articles. The principle is to emphasize the input from established sources.

When we converted the 38,256 MEADLINE articles to analyze the structural and temporal patterns of semantic predications, there is no information about either

the references they cited or how frequently they have been cited by other articles. It is possible to collect an equivalent dataset from sources such as the Web of Science and then extract semantic predications, which is in fact what we are currently working on. In the examples to follow, we construct semantic networks by selecting MEDLINE articles from this dataset with two options.

The first option is to select MEDLINE articles that are the top N articles from each year in terms of the number of semantic predications. For instance, the user can select top 50 or top 100 MEDLINE articles each year in terms of their values in the TC field, which is the number of semantic predications in each article. Semantic predications from the selected MEDLINE articles will be used to construct a semantic network of concepts and their relationships defined by predicates in associated semantic predications.

The second option is to select MEDLINE articles based on a generalized g-index (Egghe 2006). The number g is defined as the average of the first g occurrences of semantic predications per MEDLINE article. Using the g-index has an advantage over the first option. The selection of the top N in the first option is arbitrary because it does not take into account the distribution of the occurrences of semantic predications in MEDLINE articles. In contrast, the second option is based on the g-index, which provides a less arbitrary cutoff point.

Option 1: Top N MEDLINE Articles

A semantic network was generated from top 50 MEDLINE articles between 1980 and 2016 in terms of the number of semantic predications per article. The resultant network consists of 338 UMLS concepts that appeared either as the subject or the object of a semantic predication. The largest connected component (LCC) of the network contains 331 concepts, or 92% of the entire network. The modularity of the network with respect to the partition by its clusters is 0.4125, which is in the middle of the range. The average silhouette of the network is relatively low at 0.267, which means the heterogeneity of a cluster is generally high. In other words, the diversity of predications in a given cluster is relatively high.

Figure 8.10 depicts a visualization of the largest connected component of the network without applying any link filters. Each semantic predication consists of a subject, a predicate, and an object. The subject and the object are represented by concepts defined in the UMLS metathesaurus. An UMLS concept is a term that represents a group of semantically equivalent terms. Each UMLS concept has a unique identifier CUI. In the virus example, HIV is an UMLS concept (CUI = C0019682). The HIV concept is used as the representative of 101 various kinds of semantically equivalent phrases found in text. Table 8.15 shows some of the most commonly occurred terms in text. All of these terms are mapped to the UMLS concept HIV. In addition to the term HIV itself, terms such as Human immunodeficiency virus, HTLV-III, LAV, and HIV-1 are unified under the same UMLS concept HIV. There are 101 such terms identified as the subject of a

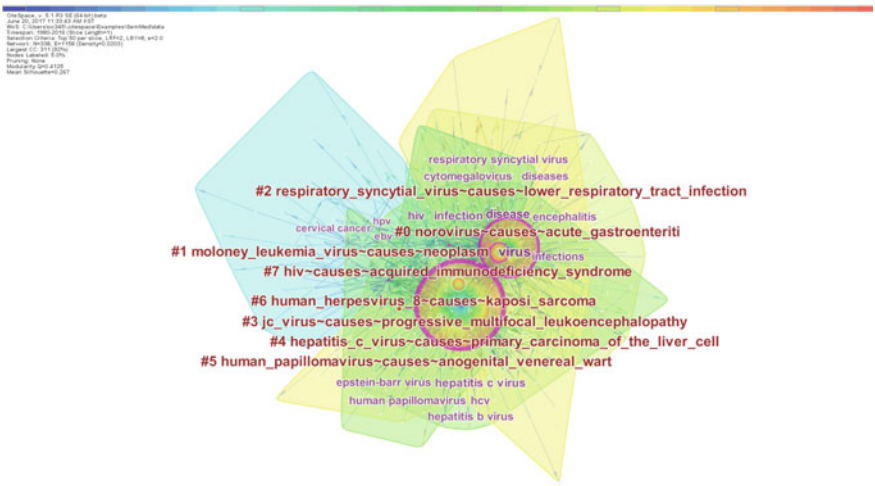

Fig. 8.10 A semantic network of 338 UMLS concepts connected by 1158 semantic predications of causality relations (1980–2016). CiteSpace: Top 50. Largest CC: 331 (92%). Q: 0.4125. S: 0.267

Table 8.15 Most frequent terms mapped to the UMLS Concept HIV (CUI = C0019682) as subjects and objects

Count	As subject	Count	As object
4892	HIV	18841	HIV
1562	Human immunodeficiency virus	4041	Human immunodeficiency virus
138	HTLV-III	460	HIV-1
98	LAV	237	HTLV-III
66	HIV-1	172	LAV
34	Human immunodeficiency viruses	132	AIDS virus
30	HIV-1LAI	90	HIVDR
23	AIDS virus	74	HIV-1IIIB
22	LAV/HTLV-III	43	TDR
19	lymphadenopathy-associated virus	41	lymphadenopathy-associated virus

predication and 268 such terms identified as the object of a predication. The mapping is done by SemRep, which we have introduced earlier.

Each node in the network is an UMLS concept, such as HIV. Two concepts are connected by corresponding predicates through semantic predications. For example, concepts of HIV and AIDS can be connected by the predicate CAUSES through the semantic predication HIV CAUSES AIDS. Individual semantic predications serve as local constraints on UMLS concepts and connect them based on their roles in semantic predications. The network therefore is a semantic network because the connections are defined by the predicates in their relationship.

CiteSpace divides the network into clusters of nodes that are tightly connected. Nodes within the same cluster appeared more often in the same predications than nodes between different clusters. Each cluster is labeled by the most representative semantic predication that is responsible for the linkage within the cluster. Cluster labels are displayed as strings of text starting with cluster IDs #0, #1, and so on. The size of a cluster is in descending order of its ID. Cluster #0 is the largest one, followed by Cluster #1. As shown in Fig. 8.10, we can see some of the nodes are labeled such as respiratory syncytial virus, hiv, and cervical cancer. Labels of the majority of the nodes in the network are not shown because they have lower frequencies than the ones that are shown. The largest cluster of UMLS concepts are labeled as #0 norovirus CAUSES acute gastroenteritis. Cluster #7 is labeled by the predication HIV CAUSES Acquired immunodeficiency syndrome.

Once we constructed a semantic network based on the semantic predications, many visual analytic functions in CiteSpace can be readily applied to the study of these predications. Figure 8.11 shows a timeline view of the network. Each line represents a cluster. Clusters are arranged in descending order from the top of the display downwards. Figure 8.12 zooms into make the fine-grain details more readable. Large circles on the left are concepts that appeared earlier on. They are connected with subsequently appeared concepts in their own clusters through the reinforcement of semantic predications. A purple rim of a circle indicates its high betweenness centrality in the network. A red ring indicates a detected period of burstness.

Figure 8.13 shows the same network such that we can identify the most frequently appeared concepts in semantic predications in Semantic MEDLINE. The size of a node represents the frequency of the concept in the virus dataset. The color of a node denotes its cluster membership. Salient concepts include virus, disease, and infection based on their size. Connections between concepts represent causal relations linked by predicates such as CAUSES.

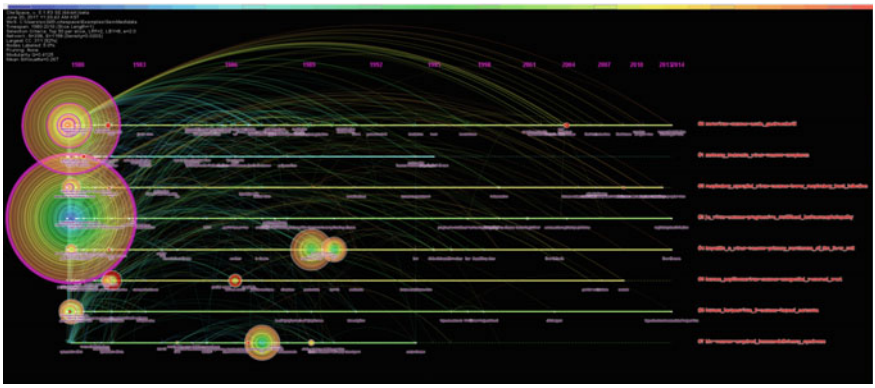

Fig. 8.11 A timeline view of the semantic predications on causality relations

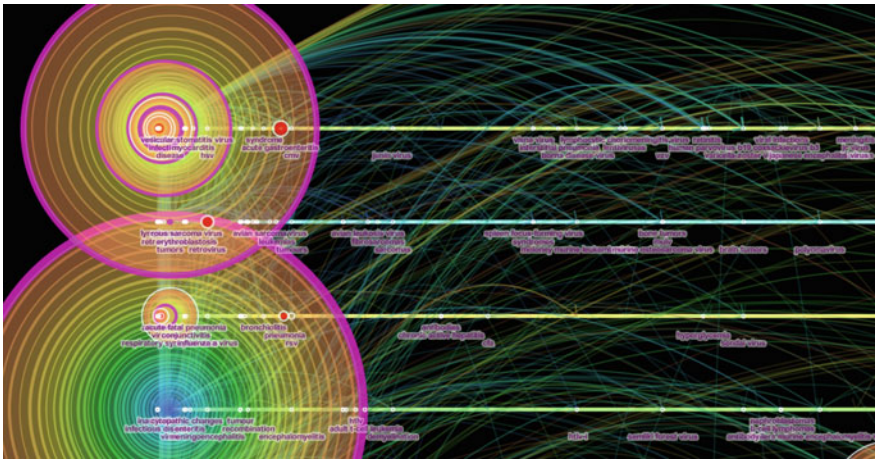

Fig. 8.12 A close-up to the timeline view of the four largest clusters of predications

Figure 8.14 shows the neighboring concepts of the concept HIV. The arrow of a link points from the subject of a predication to the object. For example, the predication HIV CAUSES Kaposis sarcoma is represented as a directed link from the concept HIV to the concept Kaposis sarcoma. Similarly, several predications are conveyed:

HIV CAUSES acquired immune deficiency syndrome.
HIV CAUSES diseases
HIV CAUSES disorder
HIV CAUSES AIDS
HIV CAUSES cytopathic effect.

The number shown on a link is the relative frequency of the particular semantic predication. The predication that HIV CAUSES Kaposis sarcoma has a value of 0.04, which is the probability of seeing the particular predication. It is based on the ratio of the number of instances of this particular predication over the number of instances of all the predications that connect the two concepts. One can also normalize the prevalence of the predication with reference to the total number of links connecting any two concepts through all semantic types.

As shown in a historical view in Fig. 8.15, the concept of HIV first appeared in 1987. Its burst was detected in 1990 and it lasted for one year, but its frequency continued to increase and peaked in 2012 involved in 94 predications that year. Figure 8.16 shows the history of the concept Virus since 1980. The concept has a period of burst that lasted for 9 years from 1980 till 1988. The concept appeared in 3481 MEDLINE articles.

From the citation history view in CiteSpace, one can look up articles that are associated with a particular concept in the semantic network of predications. In Fig. 8.17, the predication of interest is shown at point 1, cytomegalovirus CAUSES

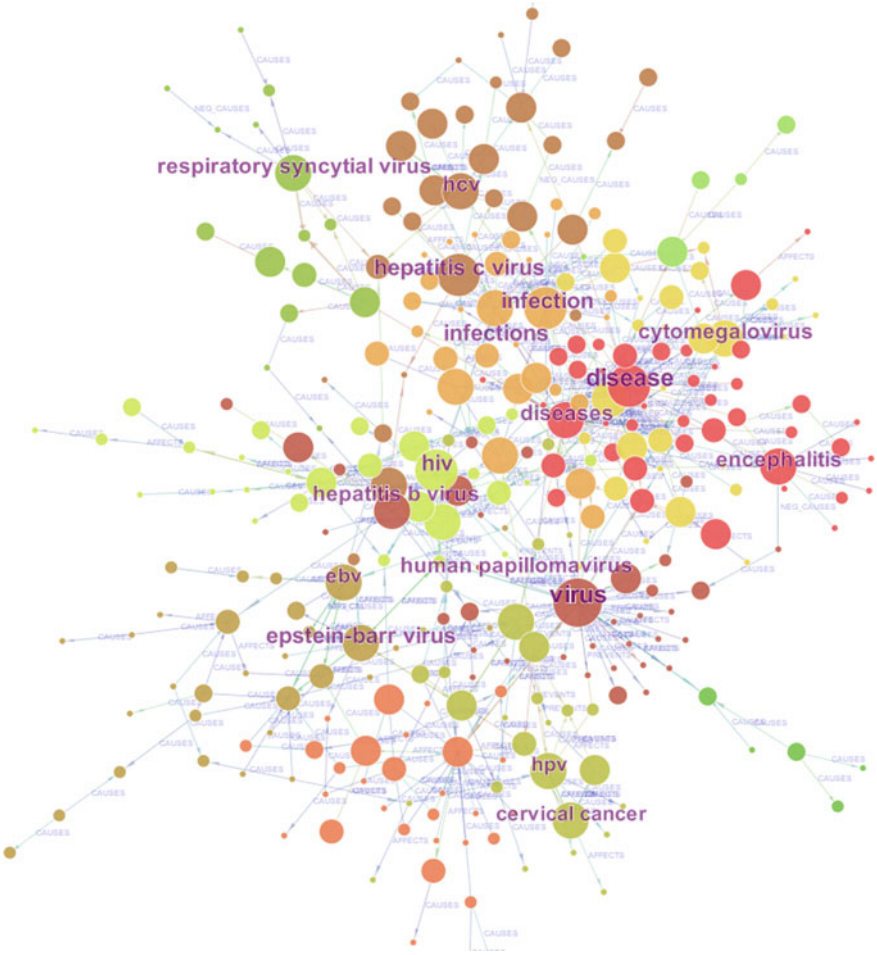

Fig. 8.13 Most frequently appeared concepts in the virus dataset

colitis. The value of 4 at point 2 indicates there are 4 MEDLINE articles containing this predication. The metadata of one of the articles is shown in the figure. In particular, the location of the predication in the abstract is underlined.

The network visualized in Fig. 8.10 is rather crowded. Clusters are overlapping with one another considerably, which affect the clarity of the view. The visualization shown in Fig. 8.18 has improved the clarity by pruning the excessive links from the network and preserving only the salient links through an algorithm called Pathfinder network scaling. The result of Pathfinder network scaling is called a Pathfinder network. Links in a Pathfinder network must meet a triangular inequality condition. Otherwise, links that fail to meet the conditions are removed from the network. In this way, the number of links is reduced while the integrity of the network is maintained by the condition.

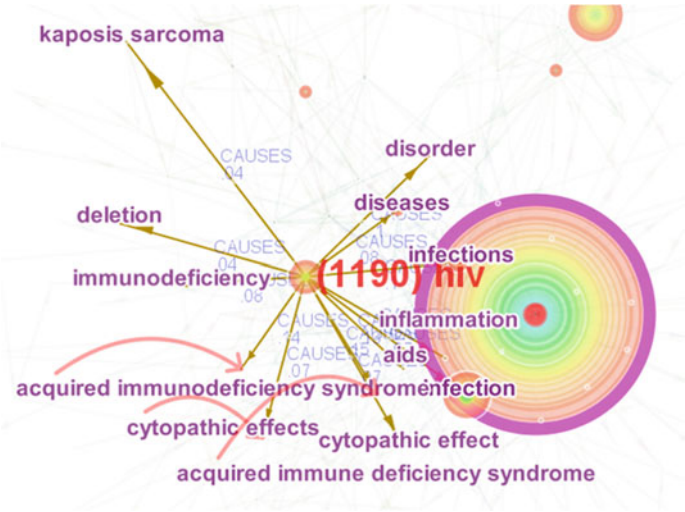

Fig. 8.14 The concept hiv and its neighbouring concepts connected through causal connections

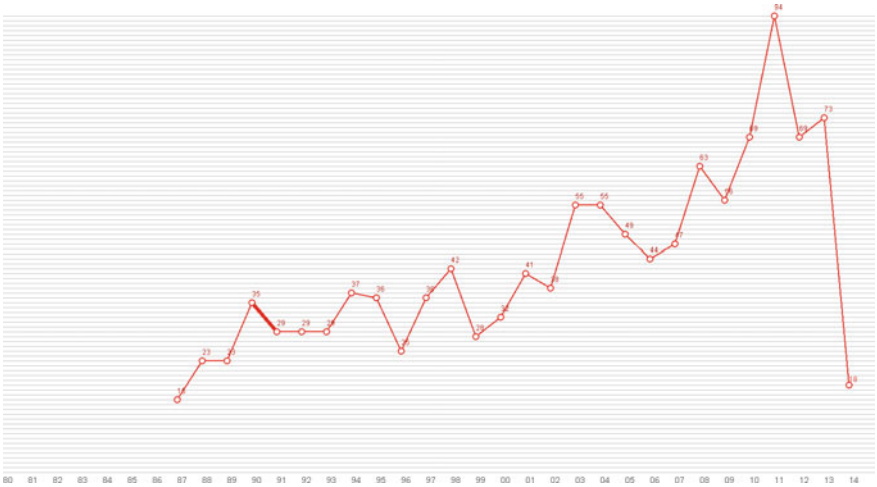

Fig. 8.15 The HIV concept has a burst of 6.7642 between 1990 and 1992. It appeared in 1190 PubMed records

The triangular inequality condition requires that the cost or weight of a direct link between nodes n_i and n_j should not be greater than the total cost of an alternative path that connects the two nodes. Otherwise, the alternative path provides more insightful connections than the direct link. Therefore, it is justifiable to eliminate the direct link from the network. In our everyday life, similar principles apply to many situations when we need to choose from multiple routes between two

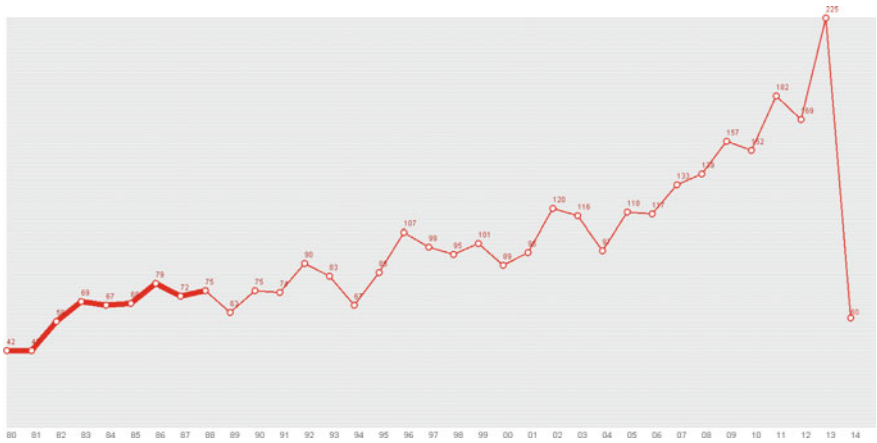

Fig. 8.16 The burstness of the concept Virus (Strength: 110.9355, duration 1980–1988). The concept appeared in 3481 PubMed records

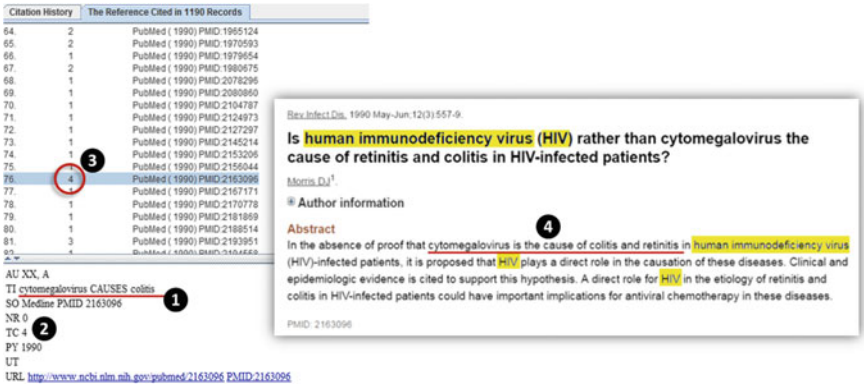

Fig. 8.17 Explore the source of a semantic predication

locations, for instance, choosing between a non-stop flight from Philadelphia to London Heathrow and a flight that makes one or two stops before London. The cost of a path could be either the door-to-door time or the price of the ticket plus the extra meals on a longer flight.

Pathfinder network scaling was originally developed by psychologists to identify major connections out of a potentially complex network. Sometimes when we compare two concepts directly, their similarity may seem low. However, once we insert the third concept in between, it may suddenly become clear how the two concepts are indeed connected through some profound ways. As soon as we see an example that can justify the closer-than-I-thought proximity, we would be more willing to revise our estimate of the similarity. The previously thought less similar concepts may appear to be more closely related.

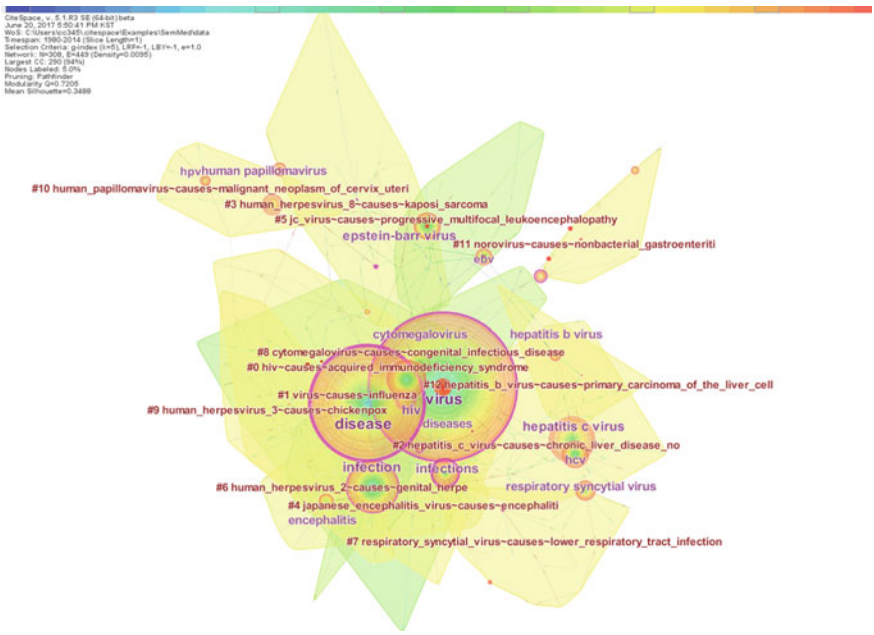

Fig. 8.18 A Pathfinder network of predications. 31 clusters labelled by LLR on predications. Node selection: g-index; Link retention: Pathfinder on time-sliced networks and the merged network

The Pathfinder network has a noticeable improvement in terms of its clarity. The largest cluster #0 is represented by the predication HIV CAUSES AIDS. The second largest one is labeled by the predication virus CAUSES influenza. The third largest one is labeled as hepatitis c virus CAUSES chronic liver disease nos. Note the nos in the label was shown as no because the stemming algorithm did not recognize NOS.

Figure 8.19 shows how the user can interact with the visualized semantic network. Upon clicking on the concept node HIV, its neighboring concepts will be highlighted while other concepts will be suppressed. An arrow from HIV points to the disease concept, representing the predication HIV CAUSES Disease. Similarly, an arrow points to the concept infections with a probability of 0.12 and a link for HIV CAUSES AIDS (0.15).

Figure 8.20 depicts a timeline view of the Pathfinder network. The nodes are selected based on their g-index. Several large clusters have high-frequency concepts, which are shown as large circles.

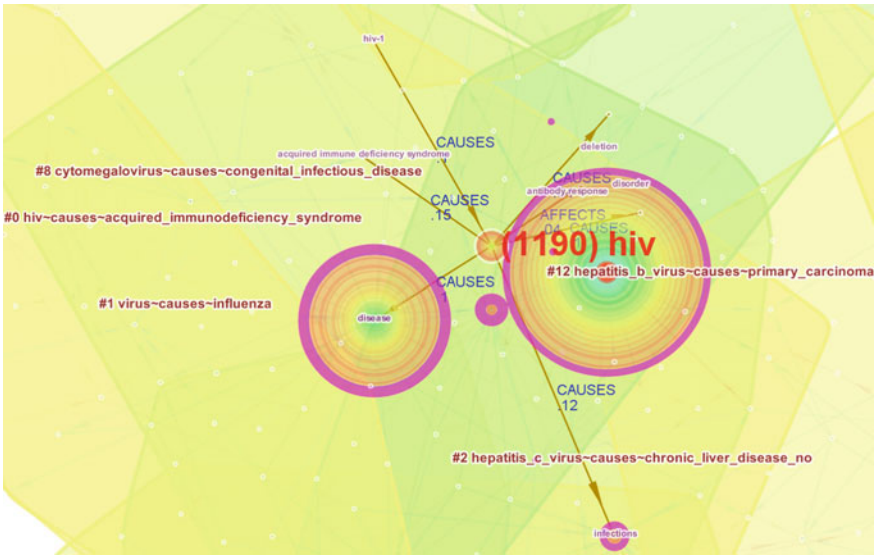

Fig. 8.19 The concept HIV and neighboring concepts. For example, HIV causes AIDS (0.15)

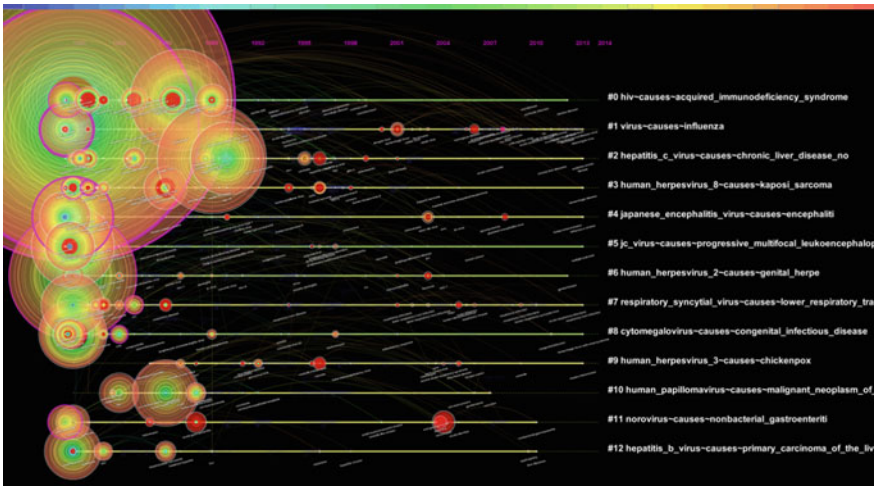

Fig. 8.20 A timeline view of the Pathfinder network. Nodes are selected by their g-index scores

Figure 8.21 shows the same timeline after the user zoomed in. The first line has several nodes with red rings. These red rings depict the durations of detected burst. The purple rims indicate concepts with high betweenness centrality scores in the network. The slightly slanted labels identify the three most frequent concepts each year in their corresponding clusters. For instance, the rightmost node on the second lowest line is labeled as human papilloma virus, which has a period of burst.

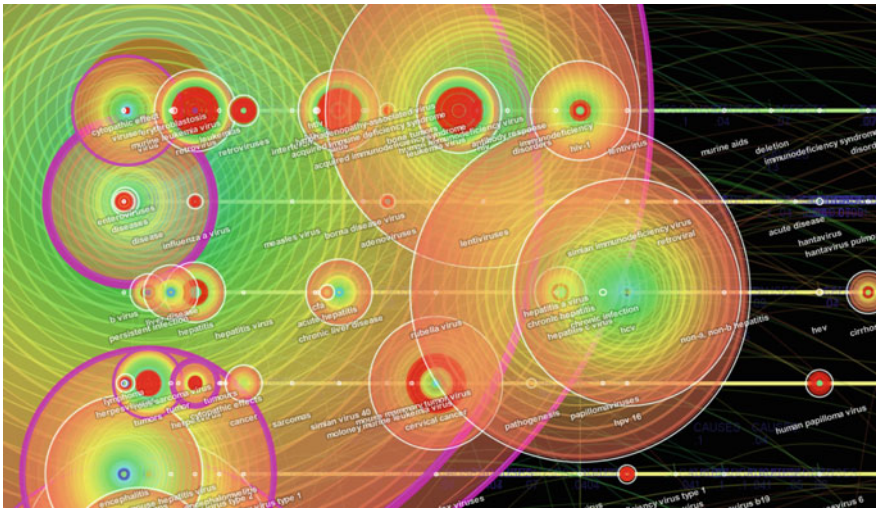

Fig. 8.21 Zooming in

Structural Variations

Analyzing semantic predications in their semantic network allows the analyst to examine novel connections added to the network. More importantly, sometimes a local structure's variation may lead to a change of the global structure. Semantic predications that have the potential to induce such global structural changes are considered important. In CiteSpace, global changes induced by local links are measured in terms of the rate of modularity change, the rate of inter-cluster linkage change, and the distribution of betweenness centrality change.

The theory of structural variation has intuitive interpretations of scientific change. Profound global changes of scientific knowledge may be caused by local changes, which serve as perturbations to a complex adaptive system. The process takes two stages. At the first stage, a novel connection needs to be made. Semantic connections, especially causality, that were previously thought unlikely or even never thought of, are proposed. Proposers are usually researchers who are visionary and creative. Novel connections that have the potential to broaden the current knowledge space are considered most valuable. The introduction of such new connections is likely to transform the state of the art of a scientific field. This stage may correspond to the first stage of Shneider's four-stage evolution model, i.e. the conceptualization stage. The key to the next stage is whether the novel connection can establish itself and attract enough followers to contribute towards the further development of the idea, including applying the original idea to a broad range of domains. This stage may partially correspond to the third stage of Shneider's model —the application stage. The structural variation functions in CiteSpace provide a concrete tool to identify the early sign of a potentially new conceptualization. It is

critical whether the process can reach the second stage, i.e. whether it can attract enough followers to keep the original idea alive.

The dashed lines in Fig. 8.22 are novel connections. An article published in 1983 (PubMed ID: 6870184) is responsible for these potentially transformative links. Adding these novel links induced the largest modularity change rate of 15.25. The global structure of the new network is significantly different from the network prior to the addition of these links. This is very useful information for the conceptualization stage. One can generate creative hypotheses that have not been considered in the scientific literature. Then the new hypotheses must be examined and attract enough researchers to make the new ideas sustainable.

In this case, the article published in 1983 (PubMed ID: 6870184) induced the largest modularity change rate of 15.25 (PubMed ID: 6870184). The dashed links represent unprecedented links connecting distinct clusters for the first time (within the scope of the dataset analyzed). Table 8.16 lists the semantic predications made by the article along with the year of their first appearance and corresponding PMIDs. In this case, these predications are not entirely new. They appeared prior to the publication of the 1983 article, but they did not meet the network modeling criteria to become part of the Pathfinder network.

Sometimes emerging patterns are more apparent if trajectories of novel links added by multiple articles are shown simultaneously. Figure 8.23 shows the trajectories of novel semantic predications made by the top 10 MEDLINE articles that are responsible for the strongest modularity change rates. Given that cluster labels are centered at the weight center of each cluster, the concentrations of the dashed lines suggest that novel predications are connecting Clusters #3, #1, #0, and #2. In particular, there are many novel inter-cluster links between Clusters #0 and #1.

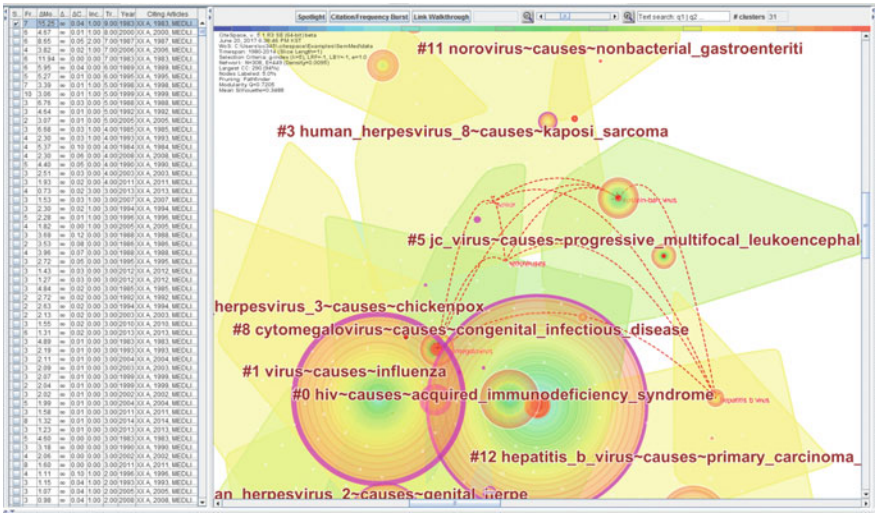

Fig. 8.22 Novel connections in dashed lines are made by a 1983 article (PubMed ID: 6870184)

Table 8.16 Semantic predications on causal relations from the 1983 article (PubMed ID: 6870184)

PID	Subject	Predicate	Object	First appear	PMID
1174925	Primary carcinoma of the liver cells	AFFECTS	Hepatitis B Virus	1978	680585
1398038	Carcinoma of Nasopharynx	AFFECTS	Herpesvirus 4, Human	1977	199059
1428074	Burkitt Lymphoma	AFFECTS	Herpesvirus 4, Human	1975	200925
1686084	Kaposi Sarcoma	AFFECTS	Cytomegalovirus	1978	212367
1840068	Retroviridae	CAUSES	Neoplasm	1979	85722
3589577	Malignant Neoplasms	AFFECTS	Human virus	1983	6870184
3589638	Carcinoma	AFFECTS	Herpesvirus 2, Human	1983	6870184

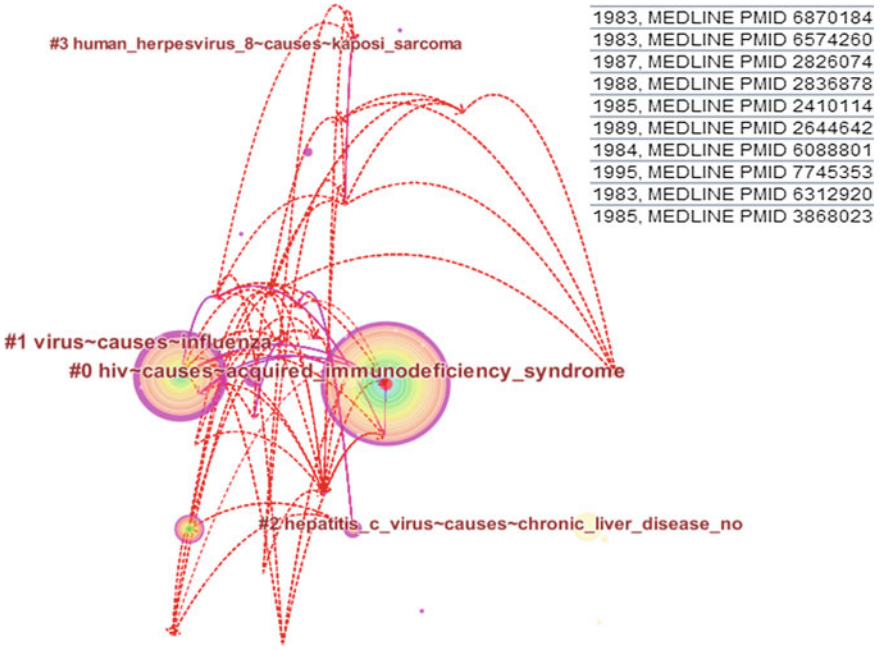

Fig. 8.23 Trajectories of novel links added by top 10 articles with the strongest modularity change rates

CiteSpace supports several ways to build an associative network. Apart from qualifying nodes based on the g-index, TopN is an alternative option. Its main advantage is its simplicity. The TopN node selection criterion selects the top N articles by citations or frequencies from each time slice to form the network. Figure 8.24 is a Pathfinder network of semantic predications between 1990 and 2014, generated with CiteSpace based on 3-year intervals and top 100 most common predications per interval.

Figure 8.25 shows the result of a structural variation analysis (Chen 2012). The semantic predications are selected from those appeared between 1990 and 2014 in 3-year intervals. Top 100 most popular semantic predications per time interval are included. The nodes of the network consists of UMLS concepts that appeared as the subject or the object of a semantic predication, such as HIV and AIDS. Connections between concept nodes are determined by semantic predications. For instance, given the predication HIV CAUSES AIDS, the concept nodes HIV and AIDS are connected in the network with a semantic link CAUSES.

The network is then divided into clusters of sub-networks based on the connectivity in the network such that concepts within the same cluster are tightly connected by semantic predications, whereas concepts between distinct clusters are loosely connected at most. In addition, Pathfinder network scaling is applied to the network, which means links that do not meet the triangle inequality condition imposed by the Pathfinder network scaling algorithm will be removed. The resultant Pathfinder network preserves the links that satisfy the triangle inequality

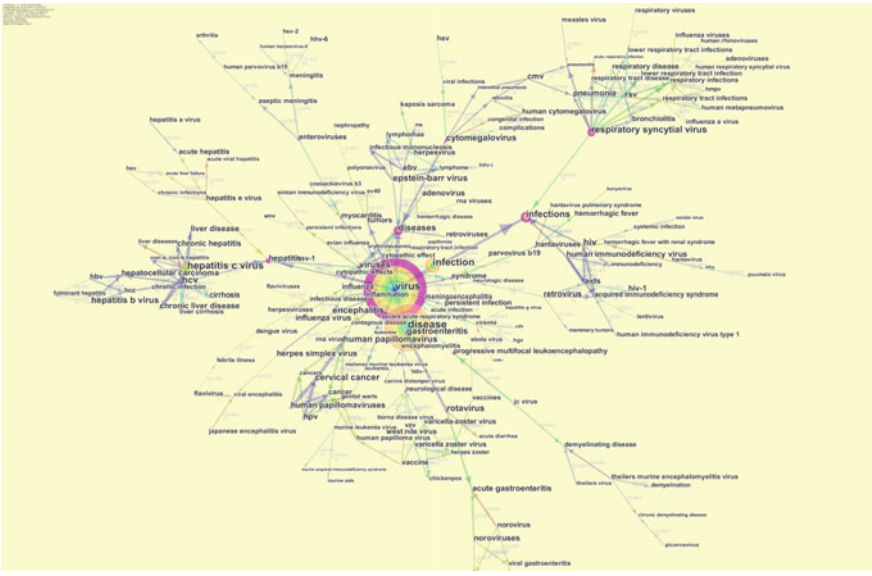

Fig. 8.24 A Pathfinder network of semantic predications generated in CiteSpace. Time slicing: 3; TopN: 100; Range: 1990–2014

Fig. 8.25 Structural variation analysis of the semantic predications (1990–2014) in CiteSpace (3-year intervals)

condition, which make the remaining structure more representative of salient semantic predications. Each cluster is labeled by the most representative concept selected by log-likelihood ratio tests in CiteSpace. For instance, Cluster #3, labeled as hiv and located near the upper right of the diagram, contains concepts such as hiv, human immunodeficiency virus, cmv, aids, retrovirus, hiv-1, and infections. Cluster #4, labeled as infectious mononucleosis and located in the mid-right area, contains concepts such as Epstein-barr virus (ebv), adenovirus, herpesvirus, and kaposis sarcoma. Dashed lines linking concepts in distinct clusters depict novel or unprecedented semantic connections at the time they appeared in MEDLINE. Such novel cross-cluster semantic connections are considered to have transformative potentials (Chen 2012).

Figure 8.26 shows novel semantic links between distinct clusters made by a MEDLINE article published in 2004 (PMID: 14766405). This article yielded a modularity change rate of 7.63, which is significantly high. It has 14 transformative

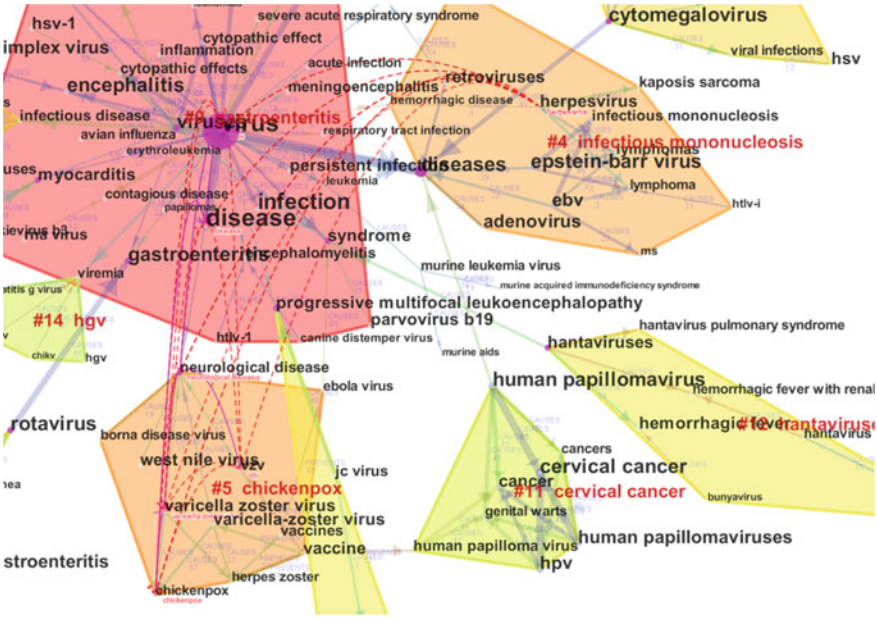

Fig. 8.26 A closer view of novel semantic links between distinct clusters made by a MEDLINE article (PMID: 14766405)

links and a centrality divergence of 0.04. These structural variation metrics indicate that since the semantic predications made by the article connect distinct clusters, there is a significant boundary-spanning potential. A boundary-spanning mechanism is considered as one of the most common types of scientific discoveries (Chen 2011).

Figure 8.27 shows annotations of the six semantic predications extracted from the MEDLINE article (PMID: 14766405). The first five predications are positive causal relations. The last one is a negation; however, the assertion is in the context of animals rather than humans, which is an important distinction that should have been preserved to avoid potential contradictions if one works on the level of extracted predications only. Furthermore, the strength of a semantic connection can be estimated based on how often the particular semantic type appears out of all the possible types connecting the two concepts. For example, the strongest link among the six relations is the first one (0.3707), followed by the third (0.1602) and the fifth (0.1454). The negation is the weakest (0.0550).

The visualized network is based on causal relations only. To obtain all semantic predications associated with the article, one can use the following query. The result is listed in Table 8.17.

```
SELECT distinct(pid), year, s_name, predicate, o_name
FROM _virus_year_text
WHERE pmid=14766405;
```

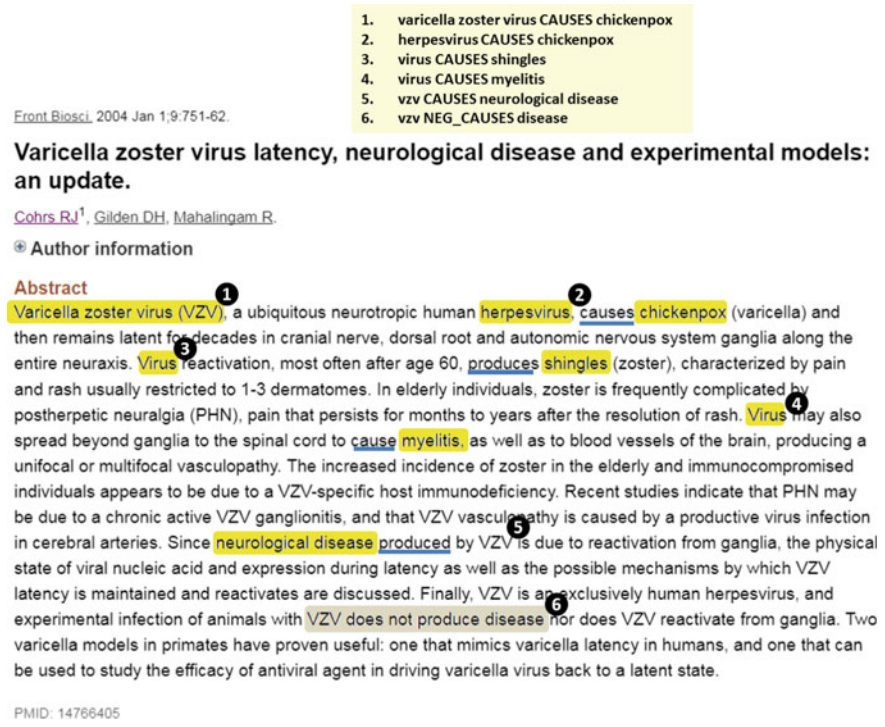

Fig. 8.27 The semantic predications extracted from the article (PMID: 14766405)

Table 8.17 All the semantic predications associated with the MEDLINE article (PMID: 14766405)

PID	First	Current	Subject	Predicate	Object
818714	1950	2004	Virus	CAUSES	Herpes zoster disease
1655451	1973	2004	Human herpesvirus 3	ISA	Herpesviridae
1968010	1977	2004	Ganglia	LOCATION_OF	Human herpesvirus 3
2680886	1986	2004	Herpesviridae	CAUSES	Chickenpox
3067201	1983	2004	Human herpesvirus 3	CAUSES	Chickenpox
3405215	1986	2004	Virus	CAUSES	Myelitis
3415496	1989	2004	Human herpesvirus 3	CAUSES	nervous system disorder
5873382	1996	2004	Posterior root of spinal nerve	LOCATION_OF	Human herpesvirus 3
10852298	2003	2004	Cranial Nerves	LOCATION_OF	Human herpesvirus 3
10852497	2004	2004	Human herpesvirus 3	NEG_CAUSES	Disease

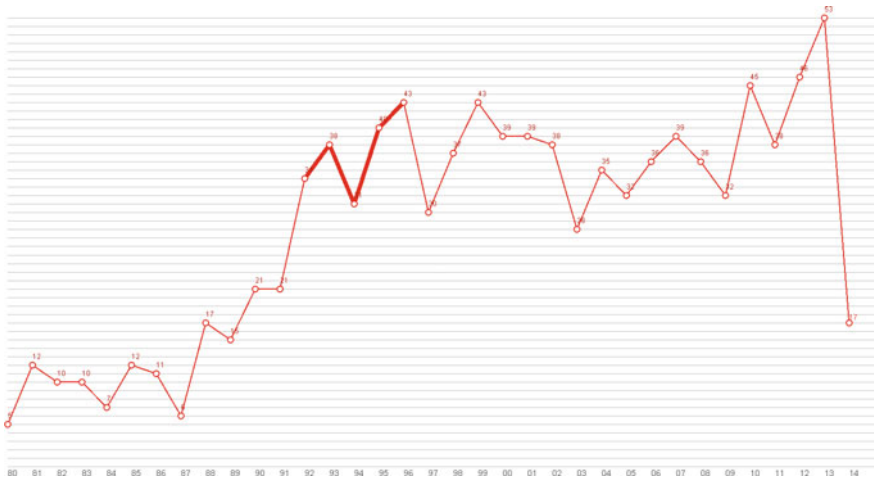

Fig. 8.28 The burstness of the concept Epstein-Barr virus between 1992 and 1995

Burst detection is a generic method. It is applicable to identify the areas of intensive interest. In this case, burst detection can be used to identify highly attractive concepts as well as semantic predications. Figure 8.28 shows the burstness of the concept Epstein-Barr Virus (EBV) in Cluster #4 of the network. Prior to the burst that started in 1992, the concept already appeared in the dataset with an average of 10 appearances each year in the first few years of the 1980s. The annual frequency jumped from 21 to 34 when its burst was detected. Since the concept is a major node in Cluster #4 given the font size of its label, the burst of the concept may indicate the emergence of Cluster #4 to a new level. Combining with the burst of semantic predications, one can explore the dynamics of research from different perspectives at multiple levels of granularity.

Option 2: MEDLINE Articles by g-Index

In addition to select MEDLINE articles based on the number of semantic predications per time slice, CiteSpace also allows the user to construct a semantic network based on the g-index. The original g-index is defined based on citations. However, given a set of predications, the user can select qualified MEDLINE articles based on the g-index of the number of semantic predications.

The following example is based on 38,256 MEDLINE articles on virus research. The relevance of each record is determined based on whether it is indexed by the MeSH term virus. Figure 8.29 shows the largest connected component of a network of co-occurring semantic predications on causal relations over a 101-year period of time (1914–2014) on virus research. In order to be included in the network, a

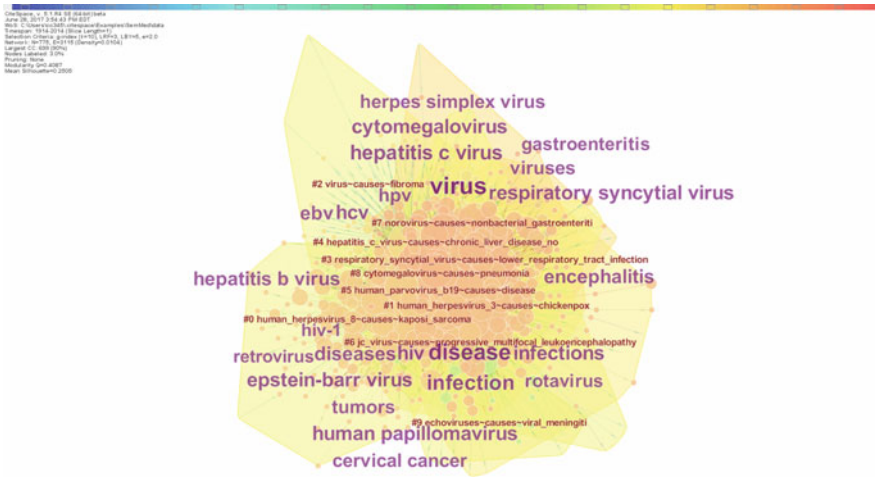

Fig. 8.29 A network of co-occurring semantic predications extracted from MEDLINE articles on virus research over 101 years (1914–2014). Node selection was based on the g-index (k = 10). Clusters of semantic predications are labelled by semantic predications with all the citing articles

semantic predication must appear twice or more in at least one year. The network contains 775 semantic predications. The largest connected component contains 699 (90%). The modularity of the network is relatively low (0.4087) as well as a low silhouette score of 0.2505, suggesting that these semantic predications are highly interrelated but the heterogeneity of each group is low.

Figure 8.30 depicts a timeline visualization of the 10 clusters contained in the largest connected component. The size of a node represents the occurrences of the corresponding semantic predication. The rings in red indicate the detected period of burst. The first 8 clusters run up all the way to 2014. The timelines of clusters #8 and #9 stopped earlier.

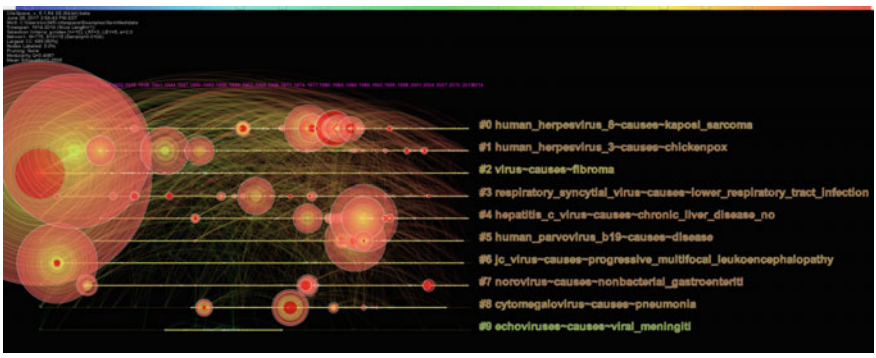

Fig. 8.30 A timeline visualization of the semantic predications on causal relations

Semantic predications in the largest cluster #0 are essentially concentrated between 1974 and 1986. The most recent predication with a burst of occurrences appeared in 1995. The cluster is labeled with the predication "human herpesvirus 8 causes Kaposi sarcoma," suggesting that the predication is the most representative for the cluster. The most representative predication for the next cluster #1 is human herpesvirus 3 causes chickenpox. Cluster #1 includes a few very popular predications in 1920s, 1940s, and 1950s. It also has a string of more recent predications with bursts.

The distribution of the circles in the timeline view indicates the activity level of each cluster (see Fig. 8.31). For example, Cluster #0 and Cluster #1 have different patterns of the distributions. Cluster #0 has predications concentrated between mid-1970s and mid-1980s, which correspond to the most active period of research in AIDS.

Figure 8.32 shows a list of 25 UMLS concepts that have a period of burst for 25 years or longer. There are many more concepts that have shorter periods of burst. These concepts may serve as the subject or the object of a semantic predication. The one with the longest period of burst is virus, which has a 74-year long lasting burst period between 1914 and 1987. Given that we are dealing with a collection of semantic predications on virus, this is hardly surprising.

The concept with the second long lasting burst is herpes virus for 67 years between 1925 and 1991. The timeline visualization shows that human herpesvirus was the subject of both Cluster #0 and Cluster #1. In Cluster #0, the most representative predication is that human herpesvirus 8 causes Kaposi sarcoma, whereas in Cluster #1, the leading predication is that human herpesvirus 3 causes chickenpox.

The third one is phage (1927–1989). The concept of phage was a focus in Cluster #5, which is labeled by the leading predication "human parvovirus b19 causes disease."

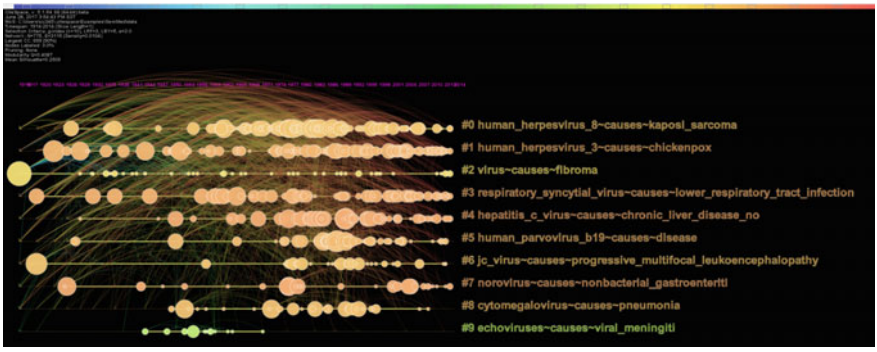

Fig. 8.31 A slightly different view of the timeline visualization with an emphasis on the distribution of predications over time

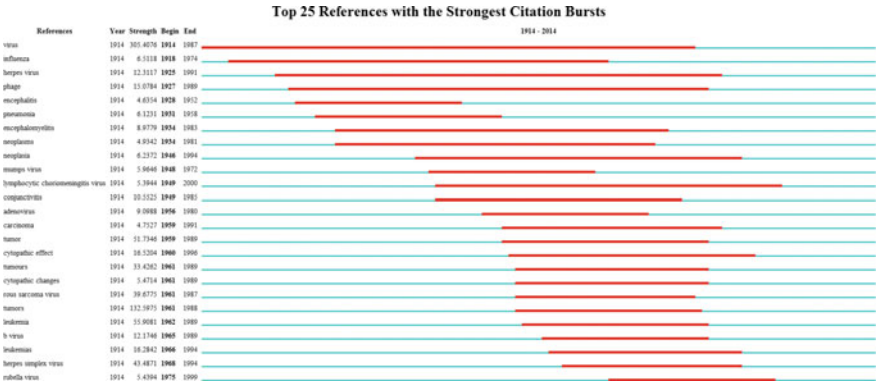

Fig. 8.32 Semantic predications with a period of burst for 25 years or longer

Influenza has the fourth longest period of burst (1918–1974). The year 1918 reminds us the 1918 flu pandemic, or the Spanish flu, which killed 50–100 million people, or 3–5% of the world's population. It was one of the deadliest natural disasters in human history.

Using the same methodology as we have applied to the study of the scientific literature of terrorism research, we generated a hierarchical structure of the semantic predications in the largest cluster (#0). The hierarchy has two branches (Fig. 8.33). The upper branch includes two major semantic predications shown in Table 8.18. We will refer them by their predication IDs in the following discussion.

Predication 7581872 on primary effusion lymphoma has two children nodes, including a branch led by predication 5292122 on Kaposi sarcoma. The former predication first appeared in the virus dataset in 1998 and first appeared in the cluster #0 in 1999 (see Fig. 8.34).

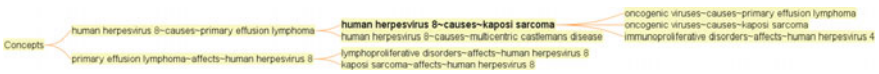

Fig. 8.33 The ontological tree of semantic predications in the largest cluster (#0)

Table 8.18 Two major semantic predications in cluster #0

Predication ID	Subject	Concept ID	Predicate	Object	Concept ID
7581872	Human herpesvirus 8	C0376526	CAUSES	Primary effusion lymphoma	C1292753
5292122	Human herpesvirus 8	C0376526	CAUSES	Kaposi sarcoma	C0036220

Fig. 8.34 The distributions of predication 7581872 in the collection of predications on virus (left) and within cluster #0 (right)

Table 8.19 presents a few examples of source sentences from which the predication 7581872 was extracted. These sentences referred to the discovery of HHV8, for example, "HHV-8 was discovered in 1994…" and the "recently identified Kaposi's sarcoma-associated herpesvirus (KSHV) and it is now "formally called human herpesvirus 8 (HHV8). Each article's PubMed ID is listed in the table.

Cluster #0 is labeled by the predication 5292122. Its distributions are shown in Fig. 8.35. It first appeared in the virus dataset in 1996 and first appeared in this cluster in 1999.

Table 8.19 Source sentences of the HHV8 and KS predication in articles published in 1996

PubMed ID	Year	Source sentences of the predication (ID: 7581872)
8627015	1996	Human herpesvirus 8 is present in the lymphoid system of healthy persons and can reactivate in the course of AIDS
8640314	1996	Human herpesvirus 8 (HHV-8, KSHV) was discovered in 1994 by means of a molecular biology approach which permitted to characterize fragments of its genomic sequence
8684008	1996	In addition, HBL-6 harbors DNA sequences of the recently identified Kaposi's sarcoma-associated herpesvirus (KSHV), now formally called human herpesvirus 8 (HHV8)
8692871	1996	Recently, DNA sequences from a novel herpesvirus, termed KS-associated herpesvirus (KSHV), or human herpesvirus 8 (HHV-8) have been identified within KS tissue from both HIV-positive and HIV-negative cases
8866603	1996	Recently, herpesvirus-like deoxyribonucleic acid (DNA) sequences, defining a new herpesvirus termed "human herpesvirus 8" (HHV8) or "Kaposi's sarcoma-associated herpesvirus" (KSHV), were detected in Kaposi's sarcoma of acquired immune deficiency syndrome (AIDS) and non-AIDS patients

Fig. 8.35 Distributions of the predication 5292122 within cluster #0 (left) and in the entire virus dataset (right)

As another example, major predications in Cluster #4 are related to hepatitis c and liver disease. As shown in Fig. 8.36, the cluster is characterized by several paths of predications, including "hepatitis c virus causes liver cirrhosis" and "hepatitis c virus causes primary carcinoma of the liver cell." The distributions of the leading predication show a steady increase since 1990 (Fig. 8.37).

Analyzing the structure and dynamics of semantic predications enables us to study the knowledge of a domain at a finer level of granularity than the conventional article-level citation or co-citation analysis. The visual analytic framework that we have developed for exploring the abstract landscape of a knowledge main provides an extensible platform for us to examine various aspects of the knowledge domain as a complex adaptive system. Each time when a new article is published, semantic predications introduced by the article serve as a source of perturbation to the current organization of semantic predications. Although perturbations act directly on local structures of the existing knowledge organization, sometimes local changes may have global and system-wide consequences. Information that can cause global changes is certainly of our interest. The following example illustrates a structural variation analysis of the high-dimensional space of relevant semantic predications on virus research. The primary goal is to demonstrate that the structural variation theory is applicable to the new level of granularity.

Structural Variations

Computing structural variation rates for the dataset is a computationally expensive. It took 14,661.465 s on a Lenovo W530 to complete the numerous but necessary comparisons required, which is just over 4 h.

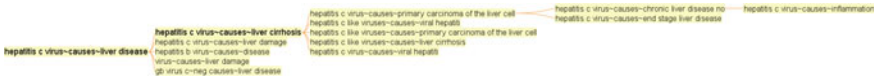

Fig. 8.36 A hierarchy of major semantic predications in Cluster #4 on relations between hepatitis c virus and the liver disease

Fig. 8.37 Distributions of the leading predication in Cluster #4 (left: the entire virus dataset and right: Cluster #4)

Figure 8.38 depicts an interaction with the visualized network of concepts connected by various semantic predications after the structural variation model is completed. The five dashed lines are novel links introduced by a 1999 MEDLINE article authored by Ellen Feigal (PMID: 9989205), entitled "AIDS-associated malignancies: research perspectives." These novel connections are derived from the following statement in the article's abstract: "The appearance in 1981 of a usually rare malignancy, Kaposi's sarcoma, in homosexual men [1] was one of the first harbingers of an epidemic caused by a retrovirus, human immunodeficiency virus (HIV), which causes the acquired immunodeficiency syndrome (AIDS)." More interestingly, these concepts belong to different clusters in the network. Linking concepts across different clusters draws our attention to this article's transformative potential. FEIGAL1999 is in fact a review article. It highlights some recent findings from the vantage point of the year 1999, including

- "discovery of a new gamma-herpes virus, human herpes virus 8 (HHV8) or Kaposi's sarcoma herpes virus (KSHV), in 1994 which led to a rapid series of investigations strengthening links of this virus in the pathogenesis of all forms of Kaposi's sarcoma (KS) [2]"
- "association of a rare type of B cell tumor called primary effusion lymphoma with HHV8 [3];"

Obviously, the two findings mentioned above are indeed the two concepts in the predication hierarchy of the largest cluster #0,—Kaposi sarcoma and primary effusion lymphoma. The FEIGAL1999 article has been cited 40 times on Google

Fig. 8.38 Structural Variation Analysis: the five dashed lines are novel links introduced by a 1999 MEDLINE article (PMID: 9989205)

Scholar. As we will see shortly, this article could be a very good starting point to understand the largest cluster of the virus research.

Table 8.20 lists the semantic predications extracted from the FEIGAL1999 article that made five transformative links, i.e. novel inter-cluster links that are unprecedented in the dataset we have examined. These semantic predications represent three types of semantic relations: ISA, PROCESS_OF, and CAUSES. For example, HIV is a retroviridae. Kaposi sarcoma is a cancer—malignant neoplasms. Retroviridae causes Kaposi sarcoma. And, HIV causes AIDS. This is a highly informative set of predications. This is part of the domain knowledge.

Kaposi sarcoma is a rare type of tumor prior to the AIDS era. It is primarily found in elderly men of Mediterranean descent and in patients on immunosuppressive therapy. In individuals with HIV positive, the incidence of Kaposi sarcoma is 75,000-fold greater and about sevenfold higher in homesexual or bisexual men than other HIV risk groups. The FEIGAL1999 review article introduced five transformative links because it focuses on infectious agents that share common etiological roles in viral infection, immune dysregulation, and cancer pathogenesis. Since the review pulls together the existing knowledge and current advances from distinct research communities such as molecular biology, immunology, virology, and anti-viral therapy, the FEIGAL1999 review is essentially serving the role of a broker of intellectual ideas originated from different disciplinary blocks. The brokerage role is likely to transform the organizational structure of the underlying domain.

It is obvious from the timeline view shown in Fig. 8.39 that dashed lines of transformative links connect concepts in different clusters. These newly added connections strengthen the tie between Cluster #0 and Cluster #5. The cross-cluster connections may be inspirational to the research community. For example, one may ask what the new relationship implies and what new discoveries would become logical. How are human herpesvirus 8 and human parvovirus b19 related? What do they have in common? If more and more articles follow up and reinforce this

Table 8.20 Semantic predications extracted from the article with five transformative links (PMID: 9989205)

PID	Subject	Predicate	Object
2383214	HIV	CAUSES	Acquired Immunodeficiency Syndrome
7348435	HIV	CAUSES	Malignant Neoplasms
2383195	HIV	ISA	Retroviridae
2050367	Kaposi Sarcoma	ISA	Malignant Neoplasms
2310015	Kaposi Sarcoma	PROCESS_OF	Male population group
1081869	Lymphoma, Non-Hodgkin's	ISA	Malignant Neoplasms
4762282	Primary central nervous system lymphoma	ISA	Malignant Neoplasms
9848528	Retroviridae	CAUSES	Kaposi Sarcoma
3723926	Retroviridae	CAUSES	Malignant Neoplasms

Fig. 8.39 FEIGAL1999 made transformative links across different clusters

pattern, one can imagine that the separation between the two clusters would diminish and eventually the two currently distinct clusters may merge into a single cluster.

Structural variation analysis in CiteSpace provides several metrics of the global changes induced by a particular article. In addition to count the number of transformative predications, one can inspect transformative changes measured by metrics such as the modularity change rate, cluster linkage change rate, and the relative entropy of the distribution of betweenness centrality. Different metrics are sensitive to different types of global structural variation. Figure 8.40 shows the footprint of

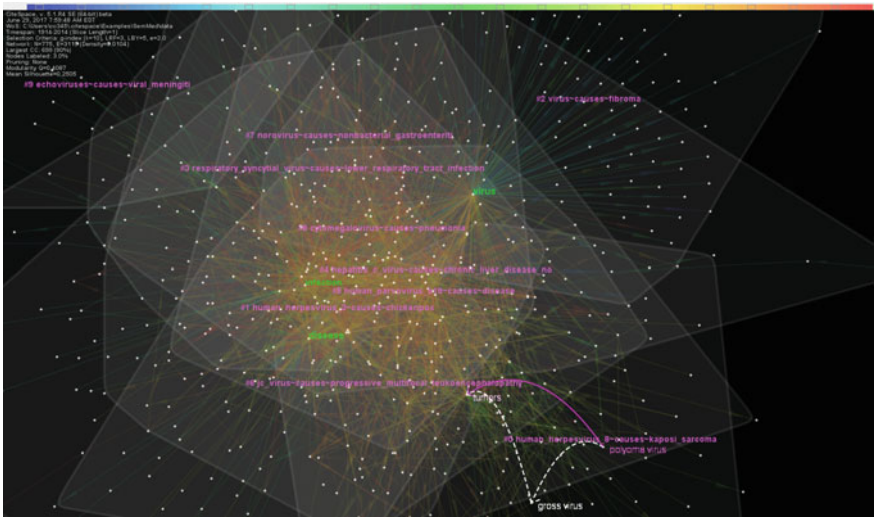

Fig. 8.40 A 1972 MEDLINE article (PMID: 4340152) added two novel predications and reinforced one existing predication. The modularity change rate induced by the article is 7.83. It also shifted the distribution of betweenness centrality scores of the nodes by a degree of 0.05

an article published in 1972 (PMID: 4340152). This article has the highest modularity change rate of 7.83. It added two novel links—one is between gross virus and tumors and the other between gross virus and polyoma virus (dashed white lines) and reinforced the existing link between polyoma virus and tumor (solid purple line).

Figure 8.41 shows the footprint of a 2001 article (PMID: 1134302), which has the largest number of incremental links of five. Unlike a transformative link, an incremental link connects concepts that belong to the same cluster. According to the structural variation theory, an article that essentially contributes incremental links is more likely to have its focus on some established research topics than articles that contribute transformative links. The article contributed semantic predications that connect concepts such as Kaposi's sarcoma and tumors, primary effusion lymphoma and tumors, kaposi's sarcoma-associated herpesvirus and tumors. All these concepts belong to the largest cluster #0. The label of the cluster is centered on the centroid of the cluster near the lower right corner of the network.

The timeline view shown in Fig. 8.42 makes it obvious—all the semantic relations contributed by the article PMID: 1134302 are within Cluster #0. The

Fig. 8.41 The footprint of a 2001 article (PMID: 1134302), which has the largest number of incremental links

Fig. 8.42 Incremental links made by article PMID: 1134302 are all within Cluster #0

Table 8.21 Semantic predications extracted from article PMID: 1134302

PID	Subject	Predicate	Object
2805744	Homologous Gene	AFFECTS	Cell cycle
5807171	Kaposi Sarcoma	AFFECTS	human herpesvirus 8
5979163	multicentric Castleman's disease	AFFECTS	human herpesvirus 8
7351621	Neoplasm	AFFECTS	human herpesvirus 8
5979188	Primary Effusion Lymphoma	AFFECTS	human herpesvirus 8
5643730	Signal Transduction	AFFECTS	Immune response
7730985	Basal Cell	PART_OF	human herpesvirus 8
1356169	Complement System Proteins	PART_OF	Virus
540196	Neoplasm	PROCESS_OF	Human

specific semantic predications extracted from the article are listed in Table 8.21, involving three types of semantic relations, namely AFFECTS, PART_OF, and PROCESS_OF. Five of the nine predications involve the concept human herpesvirus 8, which is the key concept for the largest cluster. The predication that Kaposi Sarcoma AFFECTS human herpesvirus 8 is semantically equivalent to the predication that human herpesvirus CAUSES Kaposi Sarcoma, which is the most representative predication that characterizes the cluster.

Summary

With the structural variation analysis, we can inspect the potential of an article in terms of the degree to which its semantic predications conform to the existing structure. A departure from the norm is an early sign of a potentially transformative idea. Novelty is a necessary condition for a scientific breakthrough. However, it is not a sufficient condition. A novel idea may not materialize itself for a long time. A sufficient condition of transformative research is its acceptance by the relevant research community. A revolutionary idea is not part of the knowledge of the underlying domain until fellow researchers start to pay attention to it and start to do something about it.

Citations to an article, altmetrics on social media, and the strength or length of a period of citation burst are among some of the simplest indicators of a potential impact of scientific contributions. Since these indicators do not directly reflect the structure of a domain's knowledge, they are extrinsic measures of scientific change. In contrast, metrics derived from structural variations are intrinsic measures because they directly reflect the changes of the structure of a domain's knowledge.

In next chapter, we will address a fundamental concept concerning the meta-knowledge, i.e. the knowledge of knowledge—the uncertainty of a scientific claim at a particular point of time. Take the predication that HIV causes AIDS as example. Our knowledge of what we know today may change drastically tomorrow. This is one kind of uncertain associated with our knowledge. We will discuss relevant issues in more detail in next chapter.

References

Chen C (2011) Turning points: the nature of creativity. Springer, Berlin

Chen C (2012) Predictive effects of structural variation on citation counts. J Am Soc Inform Sci Technol 63(3):431–449. doi:10.1002/asi.21694

Chen C (2017) Science mapping: a systematic review of the literature. J Data Inf Sci 2(2):1–40

Chen C, Ibekwe-SanJuan F, Hou J (2010) The structure and dynamics of co-citation clusters: a multiple-perspective co-citation analysis. J Am Soc Inf Sci Technol 61(7):1386–1409. doi:10.1002/asi.21309

Egghe L (2006) Theory and practise of the g-index. Scientometrics 69(1):131–152. doi:10.1007/s11192-006-0144-7

Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC (2012) SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics 28(23):3158–3160. doi:10.1093/bioinformatics/bts591

Kleinberg J (2002) Bursty and hierarchical structure in streams. In: Proceedings of Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp 91–101

Laudel G, Glaser J (2014) Beyond breakthrough research: epistemic properties of research and their consequences for research funding. Res Policy 43(7):1204–1216. doi:10.1016/j.respol.2014.02.006

Small H (1978) Cited documents as concept symbols. Soc Stud Sci 8(3):327–340

Wagner CS, Alexander J (2013) Evaluating transformative research programmes: a case study of the NSF Small Grants for Exploratory Research programme. Res Evaluat 22(3):187–197. doi:10.1093/reseval/rvt006