Document Semantic Representation: An Algebraic Topological Approach

Fanchao Meng

Department of Computer and Information Sciences
University of Delaware
fcmeng@udel.edu

Document Semantic Representation (DSR)

Core Task

Provide a new computational representation for a document to reflect the **semantics** of that document.

Document Semantic Representation (DSR)

Core Task

Provide a new computational representation for a document to reflect the **semantics** of that document.

Why do we need it?

Document clustering, document classification, document indexing, document comparison...

Major Contributions

DSCTP Pairwise document semantics comparison method based on **topological persistence**.

APV Vector-formed single document semantic representation based on *phrase* clustering.

GPGS Vector-formed single document semantic representation based on **graph signal processing (GSP)**.

Motivation & Related Work

Word Embedding Based Semantic Representations:

▷ **Neural Network Based:**

Word2Vec Use a word to predict its context (*skip-gram*) or the other way (*CBOW*).

NASARI *Word2Vec* trained on the UMBC corpus.

Sent2Vec Extend *Word2Vec* to sentence context.

Doc2Vec Add in a paragraph vector to *Word2Vec*.

fastText *Word2Vec* with n-grams as inputs.

▷ **Others:**

GloVe Take local context windows and co-occurrence information into a weighted least squares regression model.

LexVec Factorize positive point-wise mutual information matrix using stochastic gradient descent.

Motivation & Related Work (Cont.)

Syntactic Structure Based Semantic Representations:

Dependency Relations *UCCA*, *UDS* and *AMR*.

Constituency Relations *CCG*.

Neural Network + Syntactic Structure Semantic Representations:

Constituency Parse Tree + RNTN Take word vectors at leaves, summarize at internal nodes.

Phrase Based Semantic Representations:

Phrase2vec Take n-grams as phrases, and combine with *Word2Vec*.

Suffix Tree Take word sequences in sentences as phrases.

Motivation & Related Work (Cont.)

Goal

- Do not rely on word embeddings.
- Directly take a document as input.
- Do not use neural networks.
- Explicitly utilize syntactic structures.
- Have vector forms.

Generalized Phrases

Phrase is interesting...

We have a new idea to define **phrase**.

Generalized Phrases

Phrase is interesting...

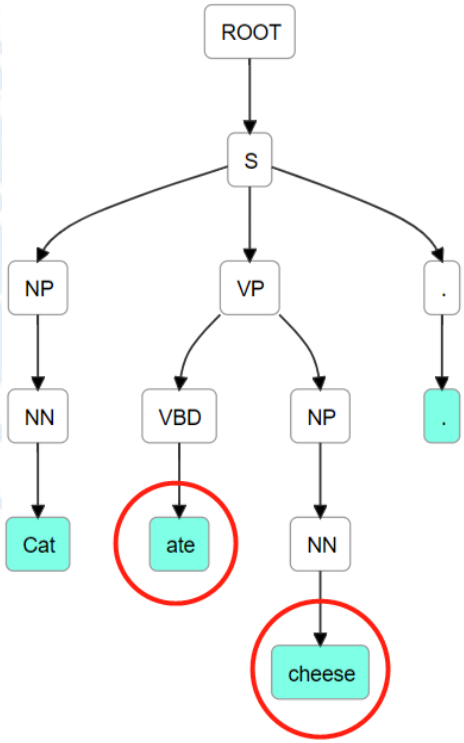
We have a new idea to define **phrase**.

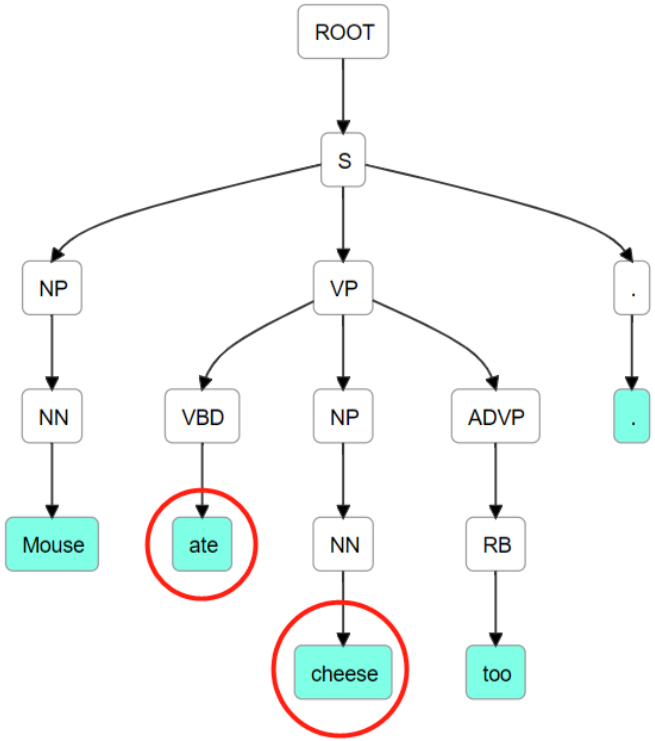
What is it?

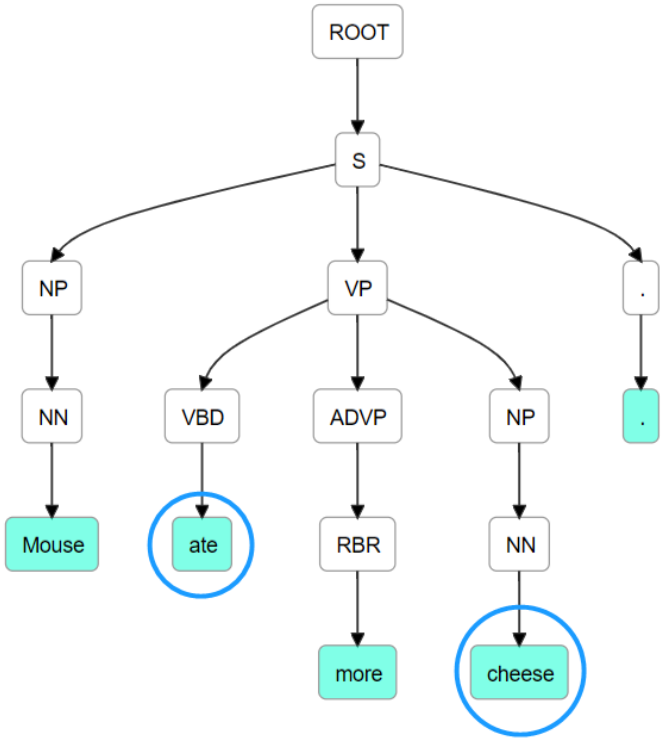
A variant of *syntactic n-grams*.

Generalized Phrases (Cont.)

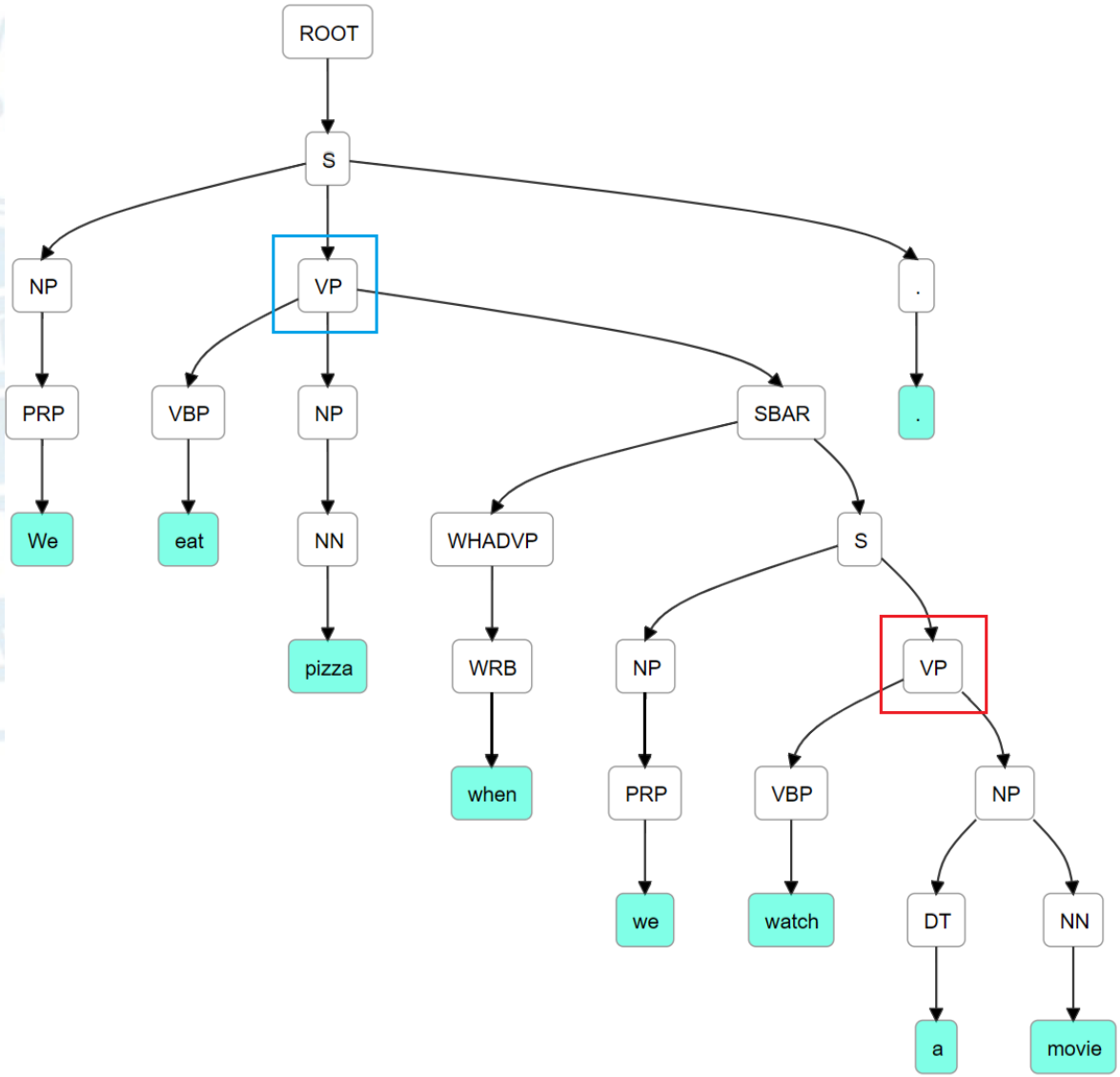
Cat **ate cheese**. Mouse **ate cheese** too. Mouse **ate** more **cheese**.





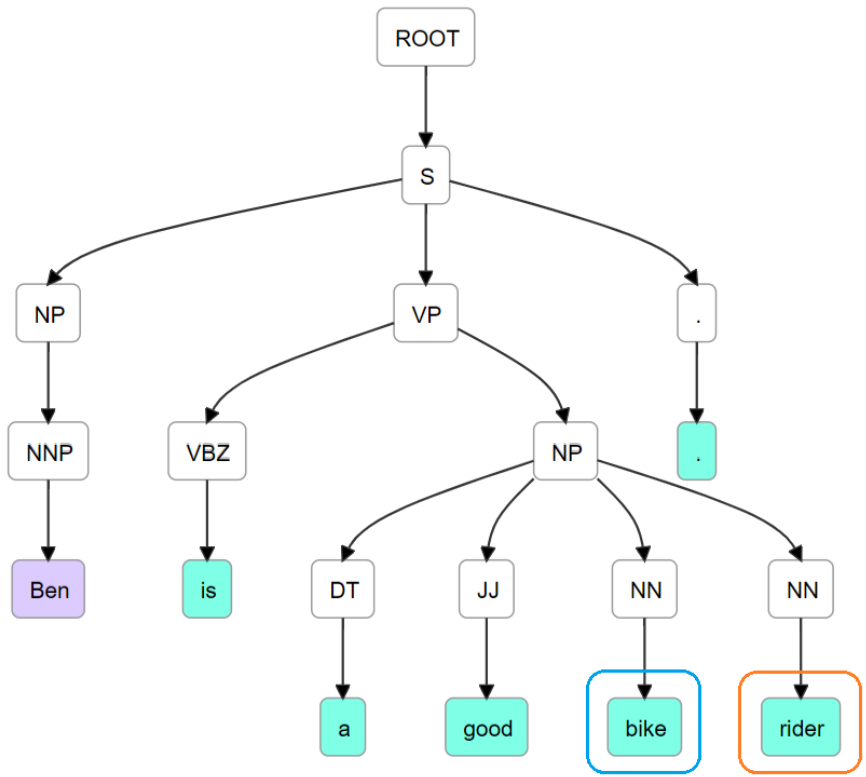


Generalized Phrases (Cont.)

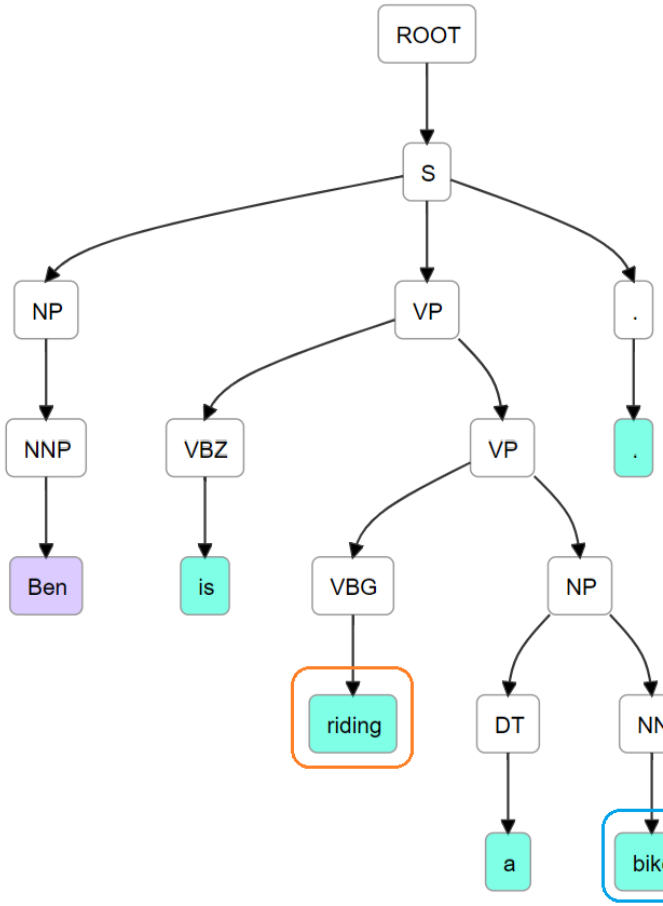


Generalized Phrases (Cont.)

Ben is a good **bike rider**.



Ben is **riding** a **bike**.



Generalized Phrases (Cont.)

Generalized Phrase (GP)

A **generalized phrase** is a minimal non-empty subtree of the constituency-based parse tree containing at most two leaves (i.e. two words). The leaves are considered as orderless. The significance of the relatedness between the leaves is computed by using the path length between them. The shorter the path length, the more significant the *GP*. If only one word is contained in a *GP*, then the path length of this *GP* is set to 1.

Generalized Phrases (Cont.)

Naturally we wonder...

Is it possible to design document semantic representations based on *GPs*?

Generalized Phrases (Cont.)

Naturally we wonder...

Is it possible to design document semantic representations based on *GPs*?

Before anything going...

We need to study if *GPs* are effective in reflecting document semantics.

DSC Problem

Document Semantics Comparison (DSC) Problem

Given:

Two documents in English consisting of one or more sentences.

Seek:

A single real value to reflect the semantic similarity between the two documents.

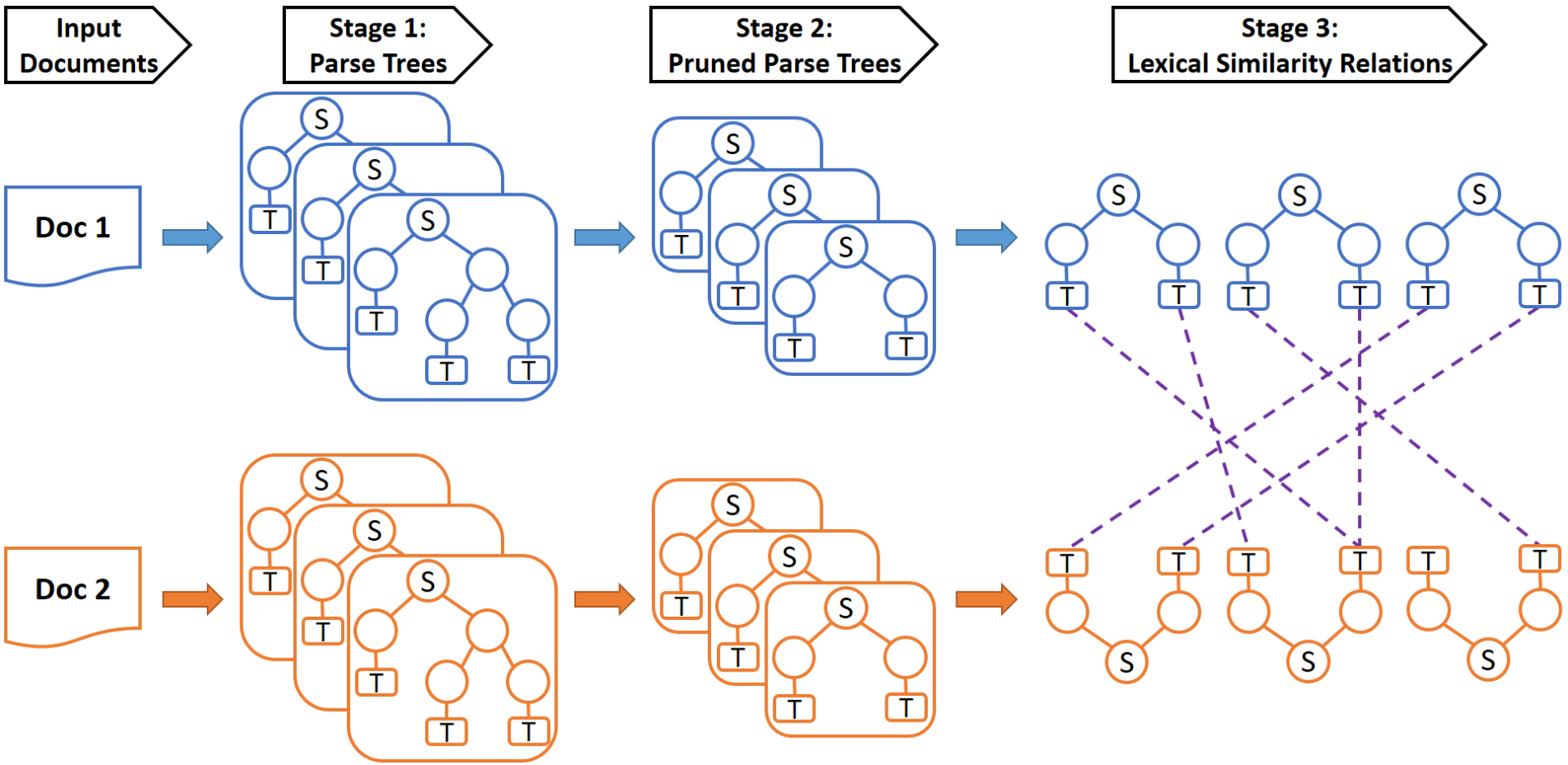
Algebraic Topology

Generally speaking...

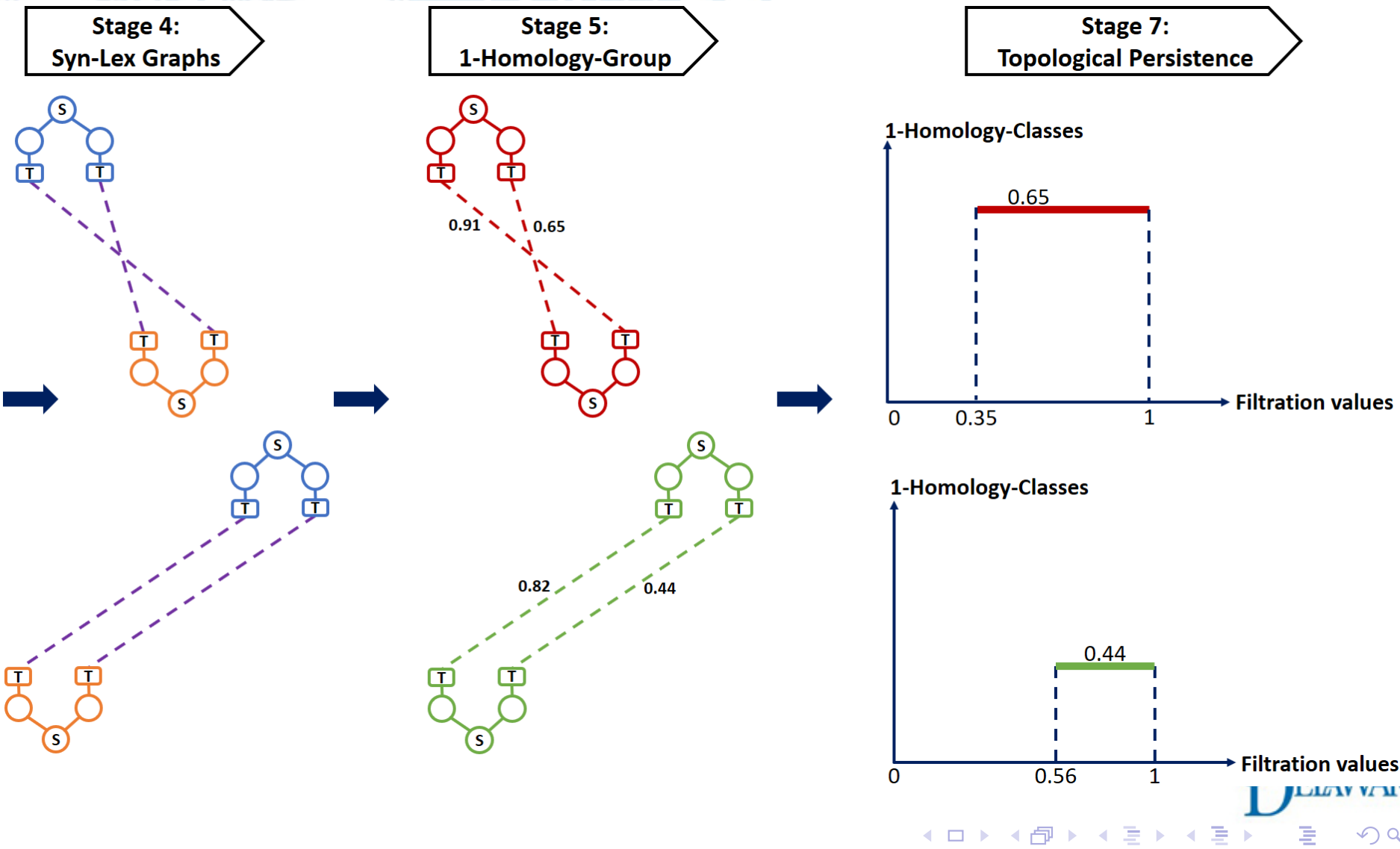
Capture "holes" in a topological space. "Holes" encode critical characteristics of the topological space.

- Abstract simplicial complex
- Homology groups
- Homology classes
- Topological persistence
- Filtration values
- Birth & death
- Lifetime
- Barcodes

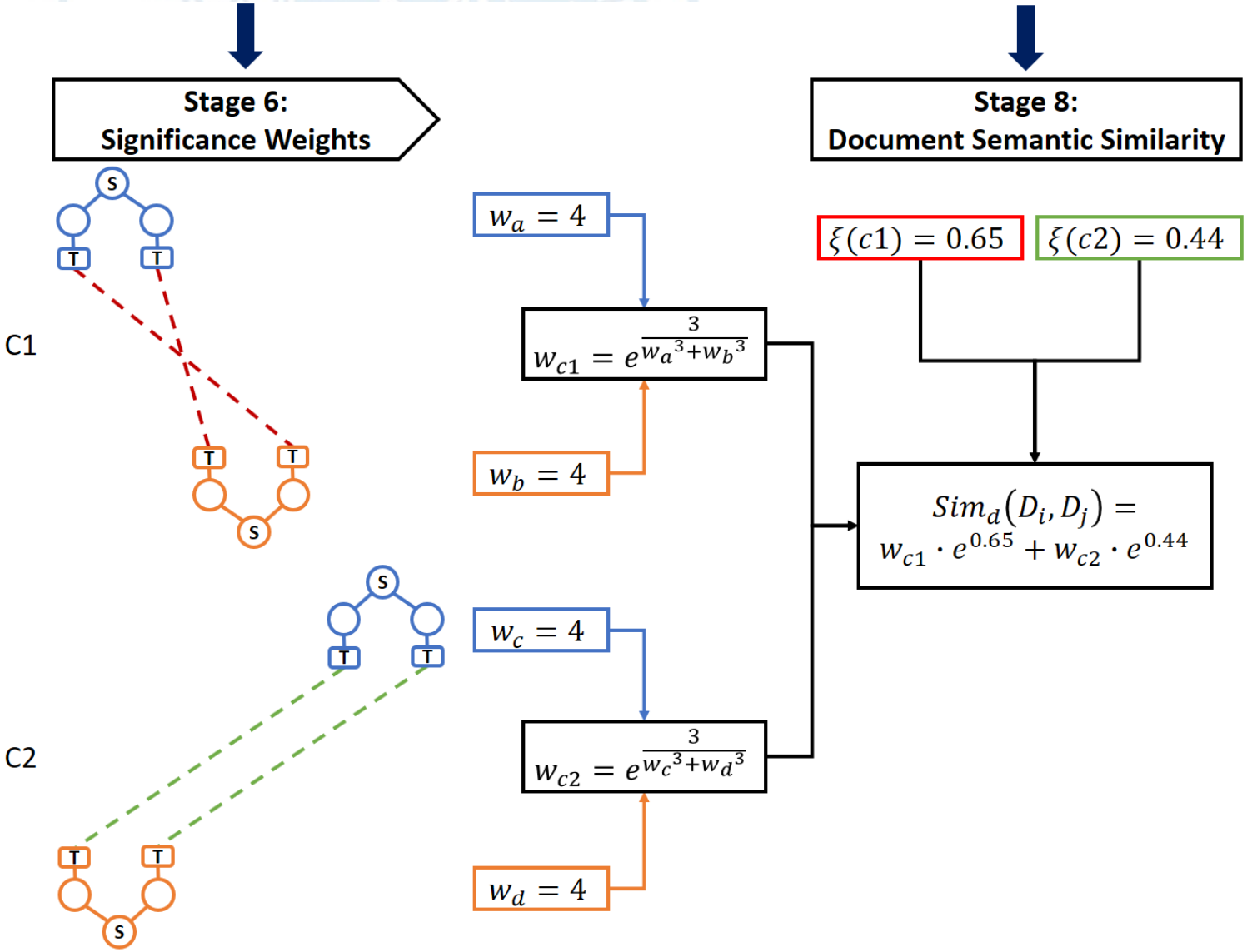
DSCTP Framework



DSCTP Framework (Cont.)



DSCTP Framework (Cont.)

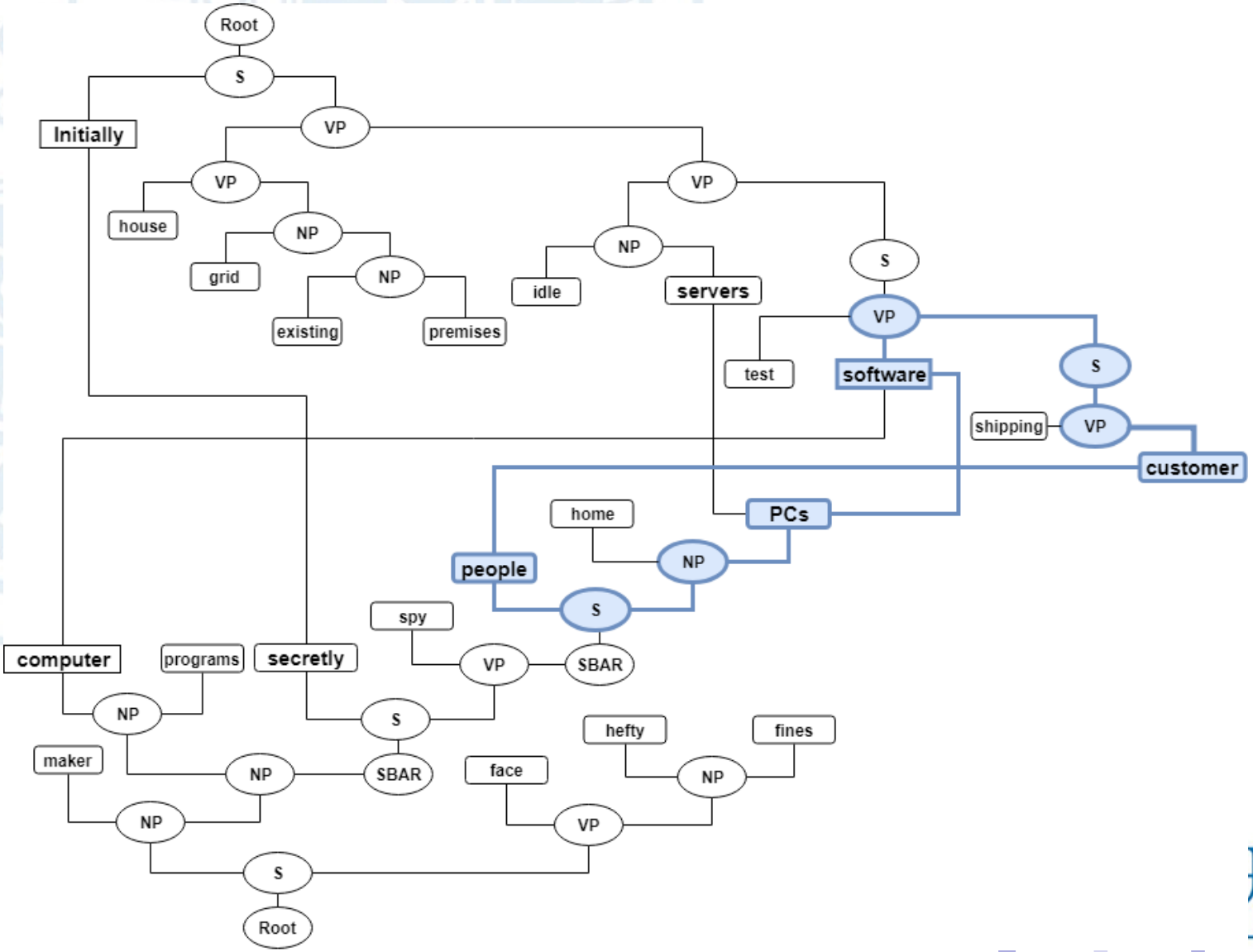


Real Example

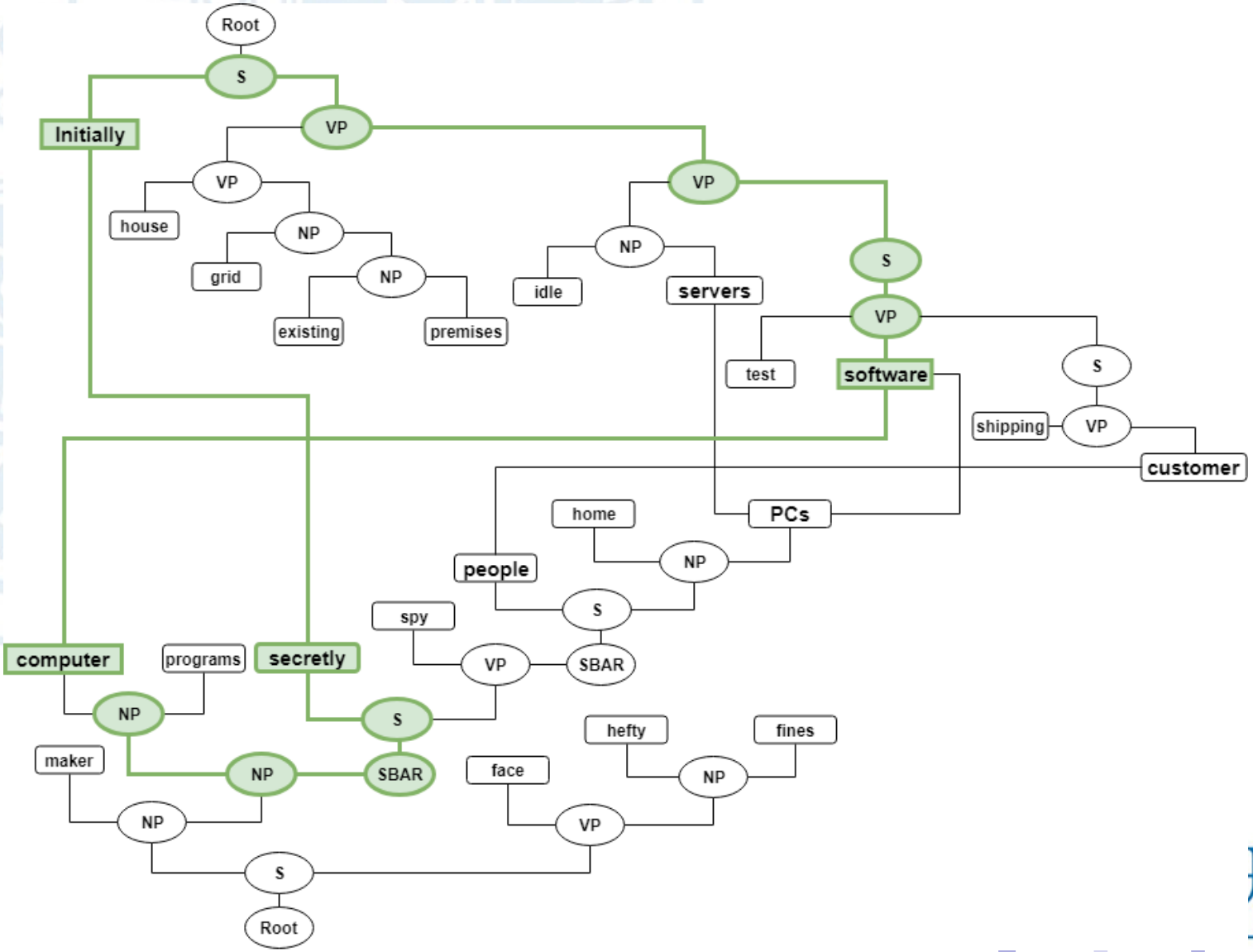
Consider two sentences...

- Initially it will house the grid in existing premises and will use idle servers to test software before shipping it to customers.
- The makers of computer programs that secretly spy on what people do with their home PCs could face hefty fines in California.

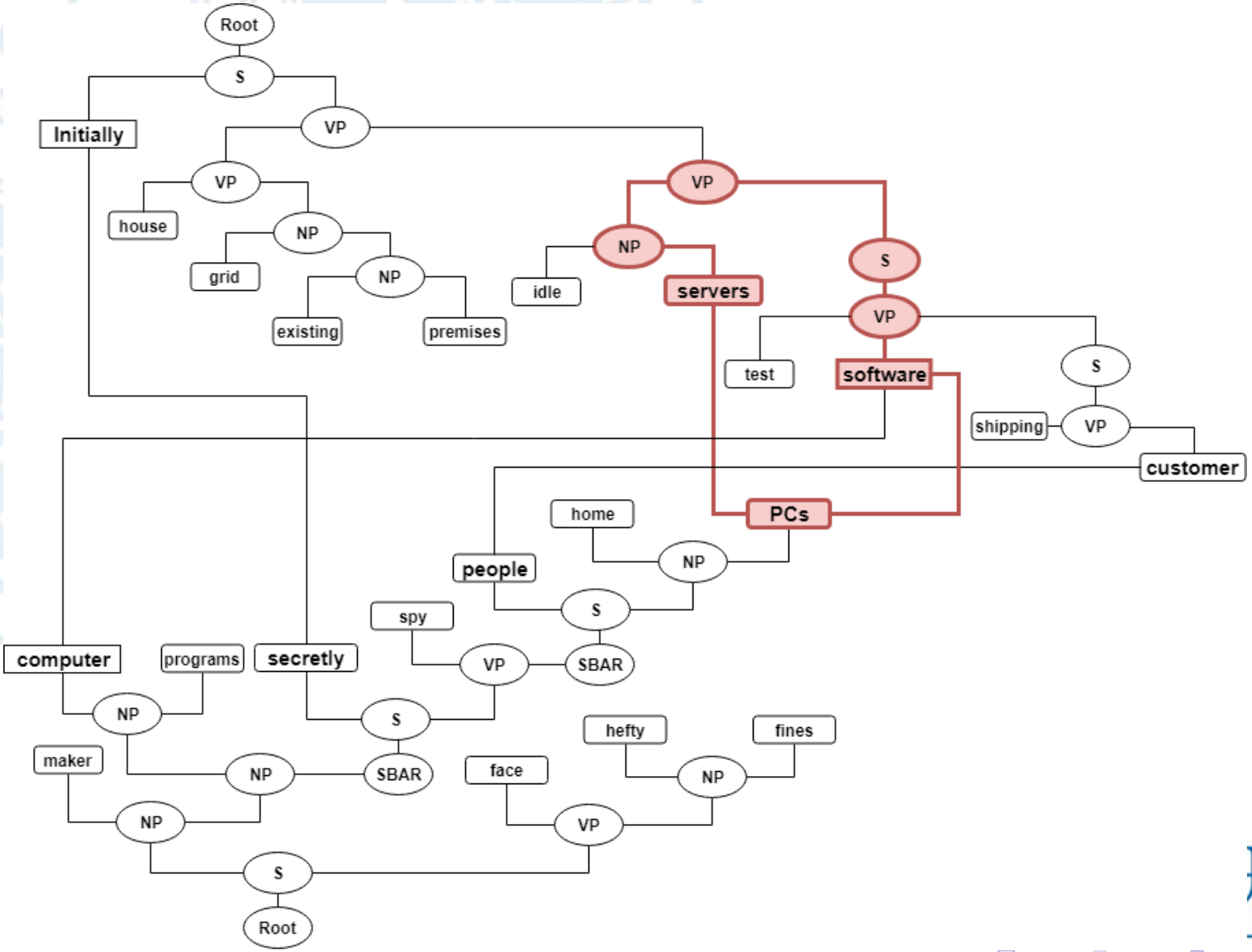
Real Example (Cont.)



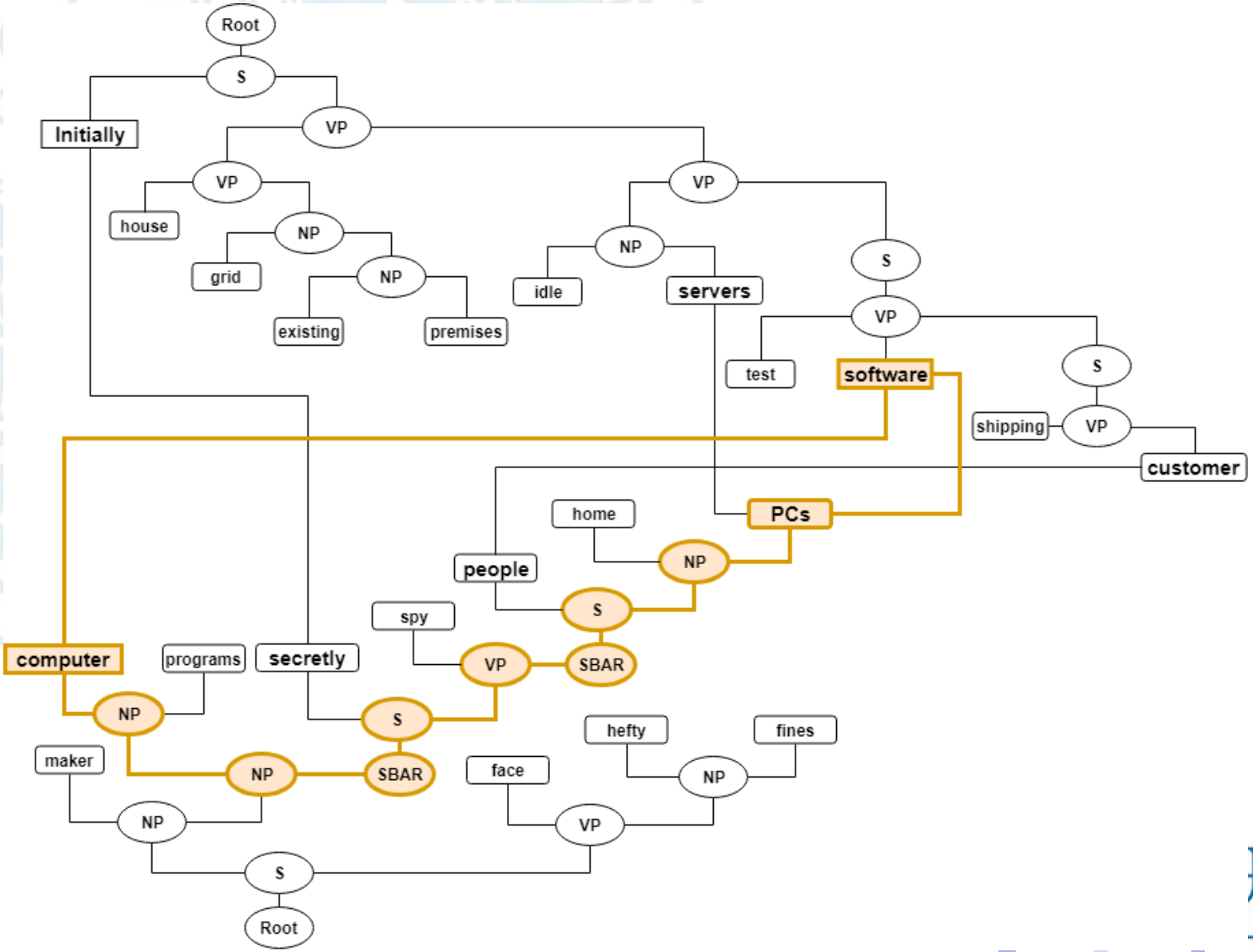
Real Example (Cont.)



Real Example (Cont.)



Real Example (Cont.)



DSCTP Experimentation

Our Methods *DSCTP+ADW*, *DSCTP+NASARI*.

Methods in Comparison *Doc2Vec*, *NASARI*, *GloVe*, *fastText*, *LexVec*, *Sent2Vec*, *WMD*.

Datasets *Lee1225*, *Lee60*, *Li30*, *STS2017* and *SICK*.

DSCTP Experimentation (Cont.)

Lee60	
Methods	**Spearman Correlation**
DSCTP-NASARI-0.3	0.73
DSCTP-NASARI-0.4	0.79
DSCTP-NASARI-0.5	0.82
DSCTP-NASARI-0.6	**0.85**
DSCTP-ADW-0.7	0.67
DSCTP-ADW-0.8	0.71
DSCTP-ADW-0.9	0.72
DSCTP-ADW-1.0	0.73
Doc2Vec	0.57
NASARI	0.79
GloVe	0.81
WMD	0.82
LexVev	0.77
fastText	0.71
Sent2Vec	0.83

DSCTP Experimentation (Cont.)

Conclusion

GPs are effective in comparing the semantics of two documents (i.e. in reflecting document semantic similarities).
This reasonably implies that *GPs* can be effective in reflecting single document semantics.

Toward GP-Based DSRs

Goal

Construct single document representations based on *GPs*.

GP Extraction

The first question to ask...

Which *GPs* should we use? All of them?

GP Extraction (Cont.)

Case 1

- "*The handheld console can* **play games**, *music and movies and goes on sale in Europe and North America next year.*"
- "*Sony has said it wanted to launch the PSP in Europe at roughly the same* **time** *as the US, but gamers will now* **fear** *that the launch has been put back.*"

GP Extraction (Cont.)

Case 2

"... the order of computation is always crucial to the functioning of the algorithm ..."

GP Extraction (Cont.)

Case 2

"... the order of computation is always crucial to the functioning of the algorithm ..."

Which topic is this document more likely about?
Computer Science or **Food**? Why?

GP Extraction (Cont.)

Case 2

"... the order of computation is always crucial to the functioning of the algorithm ..."

Which topic is this document more likely about?
Computer Science or **Food**? Why?

"... **computation** ... **algorithms** ..." $\Longrightarrow$ **Computer Science**

GP Extraction (Cont.)

Case 3

"...string ... tuning ..."

GP Extraction (Cont.)

Case 3

"...string ... tuning ..."

Which topic is this document more likely about?
Instruments or **Physics**?

GP Extraction (Cont.)

Case 3

"...string ...tuning ...standard ...guitar ..."

How about now?

GP Extraction (Cont.)

Case 3

"...string ...tuning ...standard ...guitar ..."

How about now? **Instruments**

GP Extraction (Cont.)

Case 3

"...string ... tuning ... physical ... universe ..."

How about now?

GP Extraction (Cont.)

Case 3

"...string ... tuning ... physical ... universe ..."

How about now? **Physics**

GP Extraction (Cont.)

A good *GP* extraction should satisfy...

A1. The extracted *GPs* have common-sense meanings.

A2. Capture representative individual *GPs* under a topic.

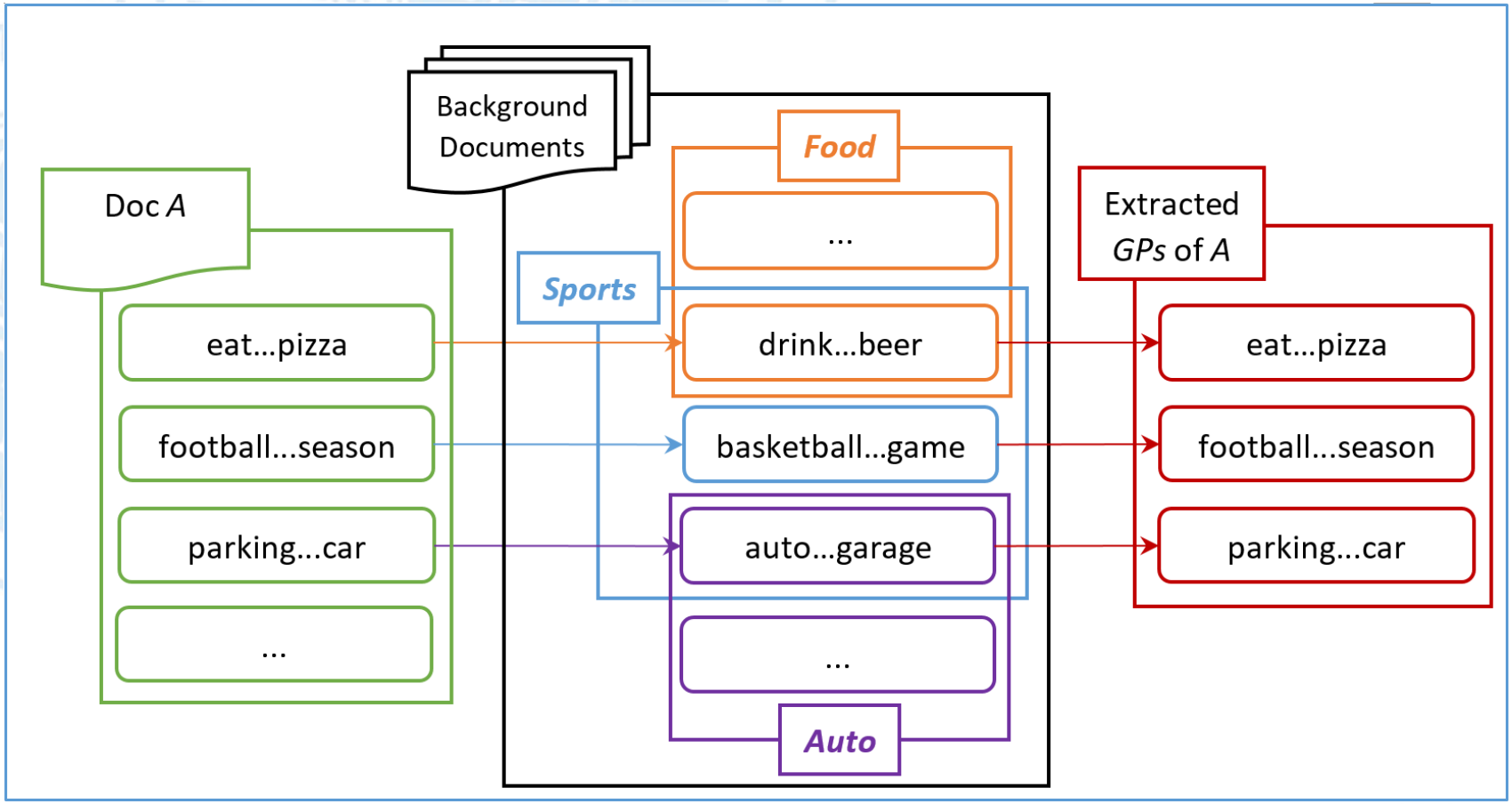
A3. Capture representative co-occurring *GPs* under a topic.

GP Extraction (Cont.)

Solution?

Background Documents + *DSCTP*

GP Extraction (Cont.)



APV Construction

Toward a Vector-Formed Representation

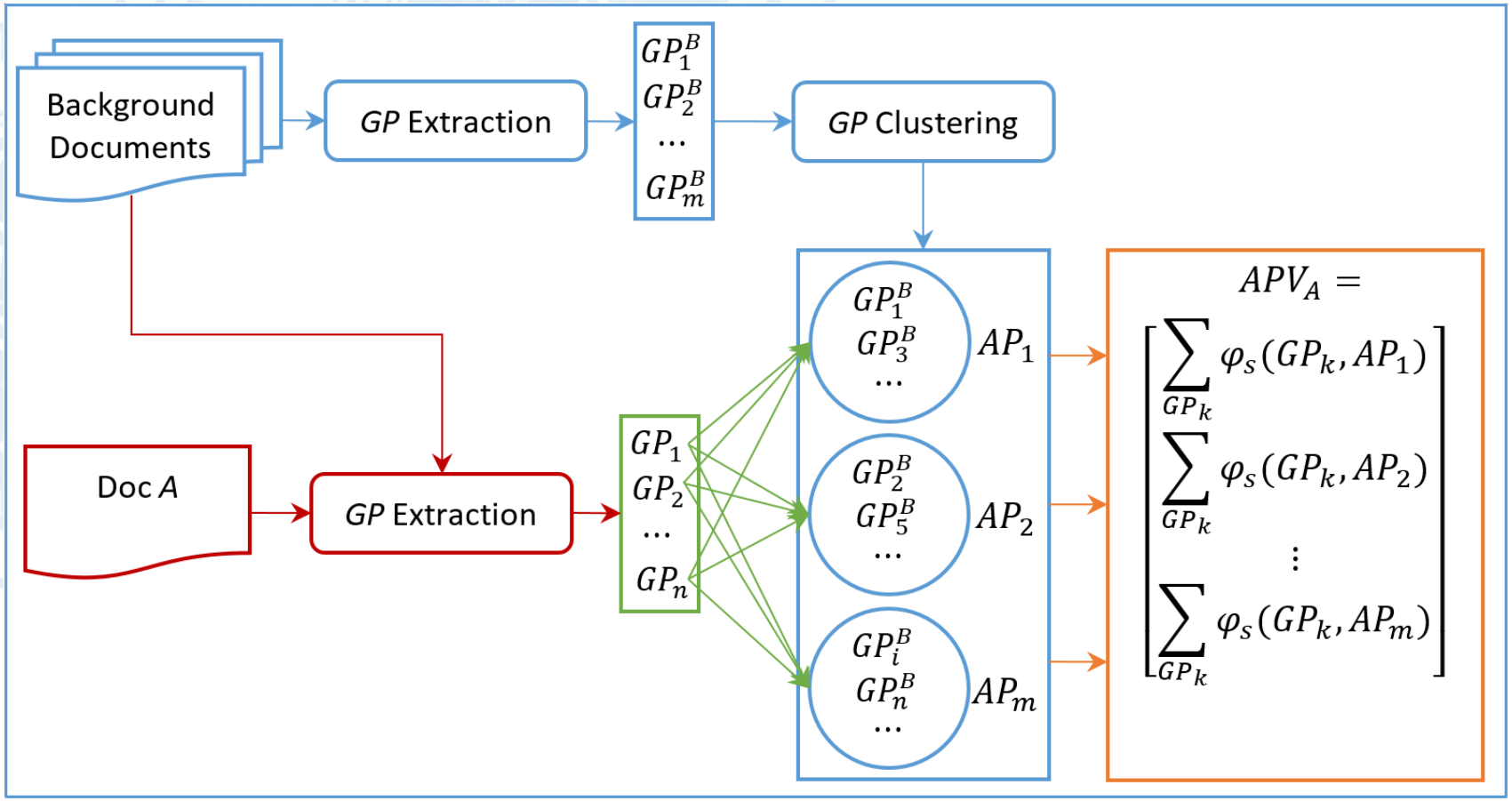
Step 1 *GP* clustering $\Longrightarrow$ Vector dimensions

Spectral clustering over *GPs* extracted from background documents.

Step 2 *GP* classification $\Longrightarrow$ Vector values

Computed based on *GP* similarities and *GP* path lengths

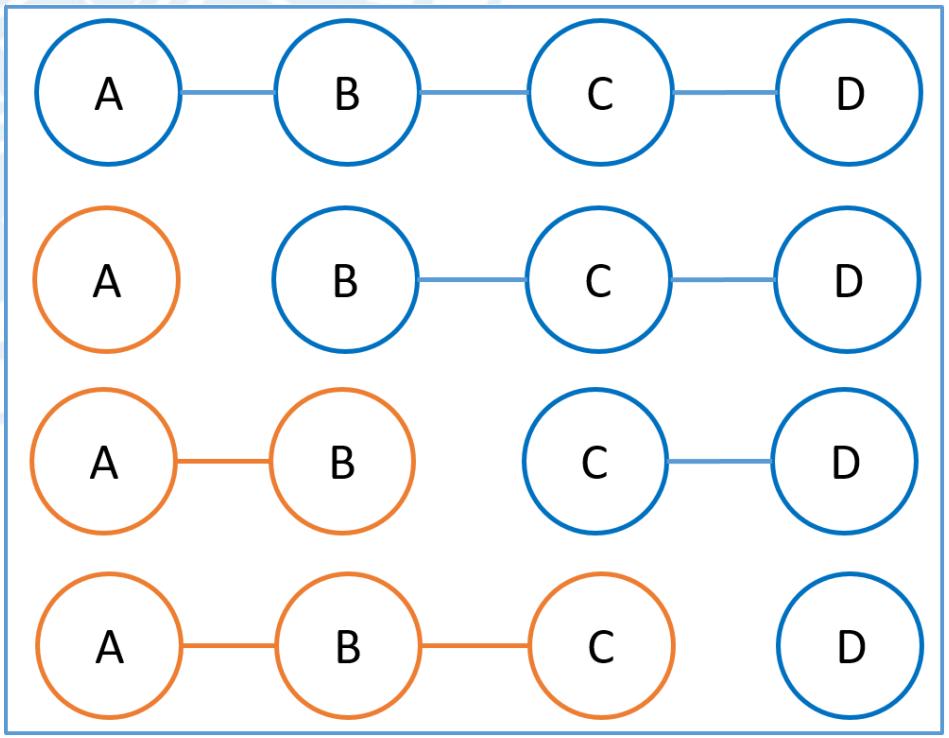
APV Construction (Cont.)



Issues

Issue #1: Confused Clustering

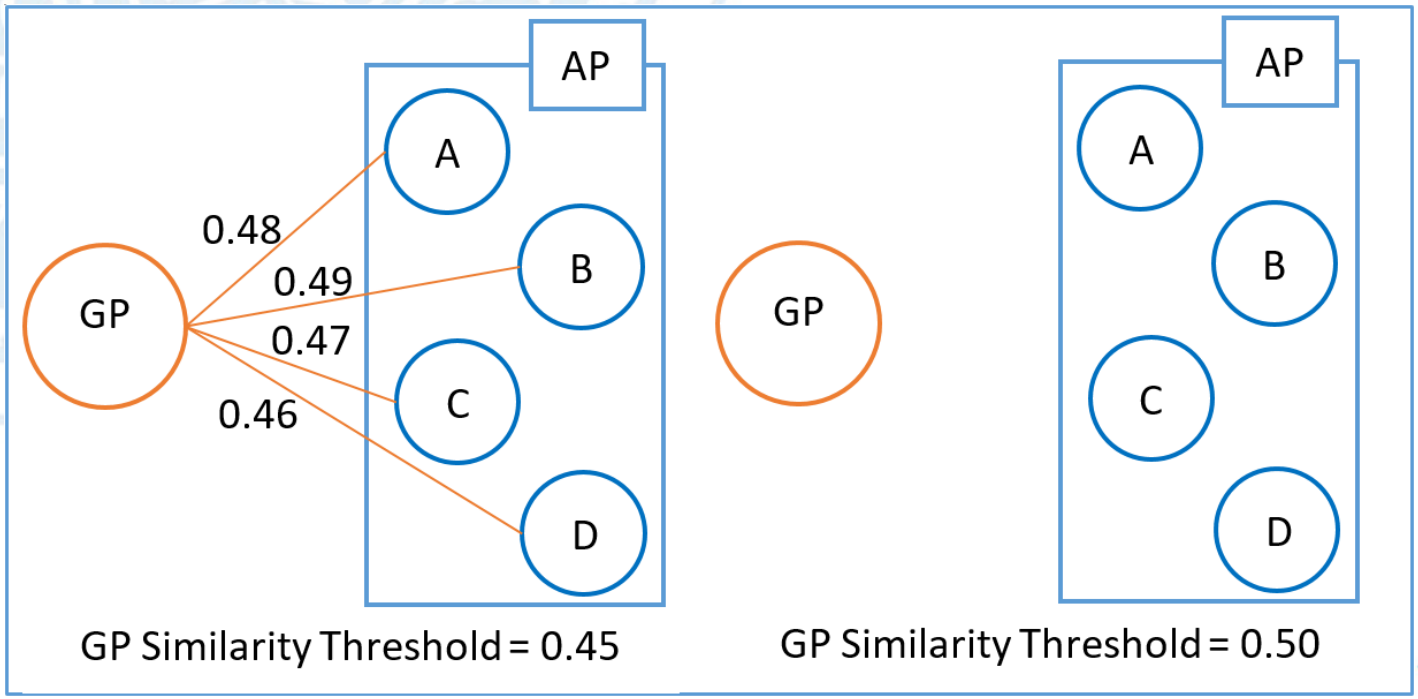
Which one is the best?



Issues (Cont.)

Issue #2: Slashed Classification

You lose a bit, you lose them all.



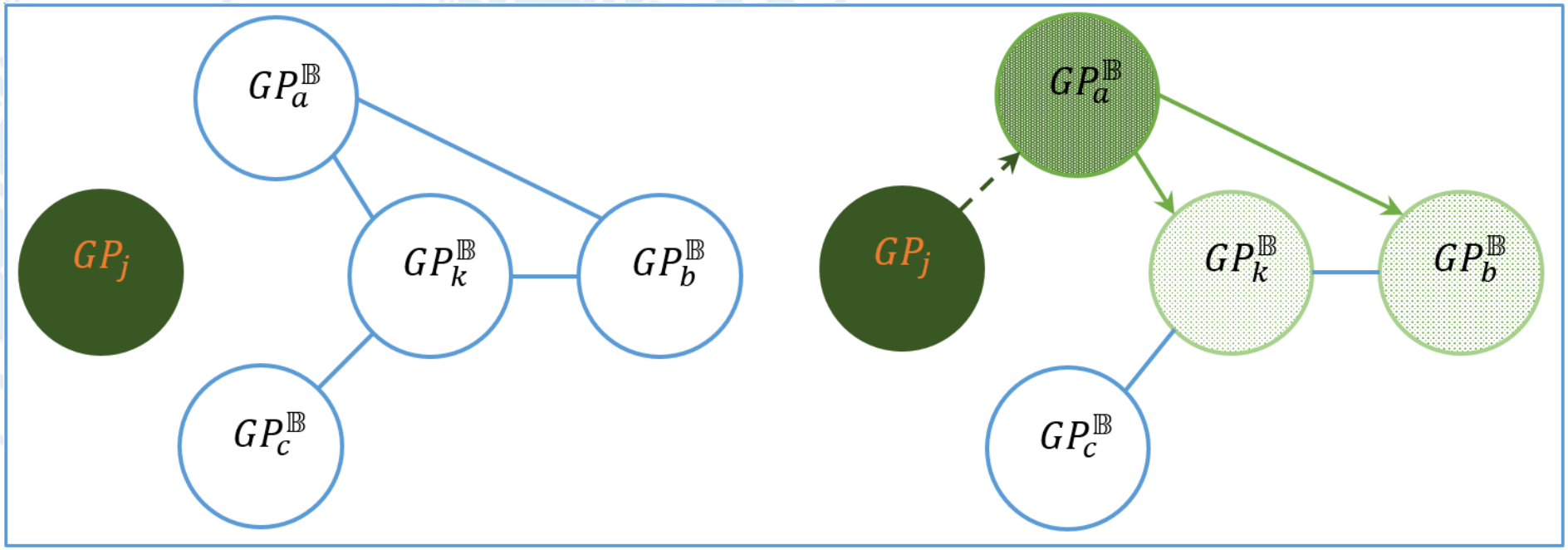
GPGS To The Rescue

Ideas

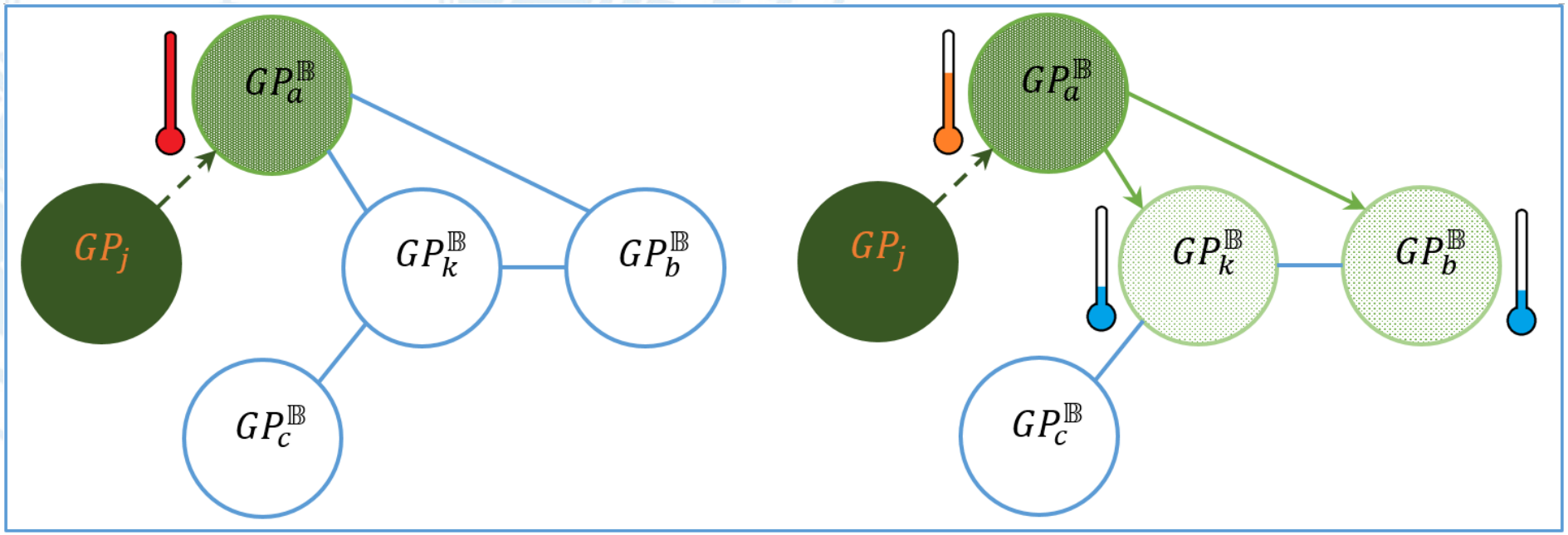
Issue #1 Do not cluster *GPs*. $\Longrightarrow$ Vector dimensions are *GPs*.

Issue #2 If X is similar to Y, X is probably similar to Y's neighbors to some extent. $\Longrightarrow$ **Induced** vector values.

GPGS To The Rescue (Cont.)



Graph Signal Processing (GSP) (Cont.)



HDC Problem

Hard Document Clustering (HDC) Problem

Given:

A set $\mathbb{D} = \{D_1, \ldots, D_n\}$ of documents.

Seek:

A partitioning of $\mathbb{D}$, where the documents in each partition are more semantically similar to the ones in the same partition than to others.

Experimental Settings

Our Methods *DSCTP*, *APV-BRSC*+Cosine, *APV-TRSC*+Cosine, *GPGS*+Cosine.

Methods in Comparison *Doc2Vec*+Cosine, *NASARI*+Cosine, *Sent2Vec*+Cosine, *LexVec*+Cosine, *fastText*+Cosine and *GloVe*+Cosine.

Clustering Method Spectral clustering

Datasets *20News-M5*, *20News-C10*, *Reuters-M7*, *BBC-M5*.

Experimental Results

20News-M5 K=5										
	DSCTP	**APV-BRSC**	**APV-TRSC**	**GPGS**	**Doc2Vec**	**NASARI**	**fastText**	**Sent2Vec**	**LexVec**	**GloVe**
ARI	**0.90**	0.84	0.87	0.88	0.86	0.79	0.64	0.56	0.85	0.80
NMI	**0.91**	0.86	0.87	0.87	0.84	0.81	0.68	0.67	0.85	0.83
FMI	**0.92**	0.87	0.90	0.90	0.89	0.83	0.71	0.67	0.88	0.84

20News-M5 K=5,6,7,8,9,10										
	DSCTP	**APV-BRSC**	**APV-TRSC**	**GPGS**	**Doc2vec**	**NASARI**	**fastText**	**Sent2Vec**	**LexVec**	**GloVe**
ARI	**0.90**	0.84	0.87	0.88	0.86	0.79	0.77	0.72	0.86	0.82
ARI-K	5	5	5	5	5	5	6	7	6	6
NMI	**0.91**	0.86	0.87	0.87	0.84	0.81	0.77	0.74	0.88	0.83
NMI-K	5	5	5	5	5	5	6	7	6	5
FMI	**0.92**	0.87	0.90	0.90	0.89	0.83	0.81	0.77	0.88	0.86
FMI-K	5	5	5	5	5	5	6	7	5	6

Conclusion

APV and *GPGS* are effective *DSRs* based on *GPs*. They are vector-formed without relying on word embedding models or neural networks, directly taking a document as input, and explicitly utilize syntactic structures.

Future Work

Social Dynamics *DSCTP*, *APV* and *GPGS* will be utilized to address the *message propagation* problem, which is a cutting-edge interdisciplinary study across social dynamics and natural language processing. This study is a part of the *SocialSim* project funded by *DARPA*. The presenter of this talk have participated in this project since 2018 summer, and will continue this work in his post-doc period.

Improve GPGS The average number of dimension of *GPGS* is high (typically higher than 1000). Dimensionality reduction may be needed, though this reduction may not be linear.

The End

Topologist's Nightmare

Figure 1: Discovered by Fanchao at *TJ Maxx* on March 17th, 2019.