

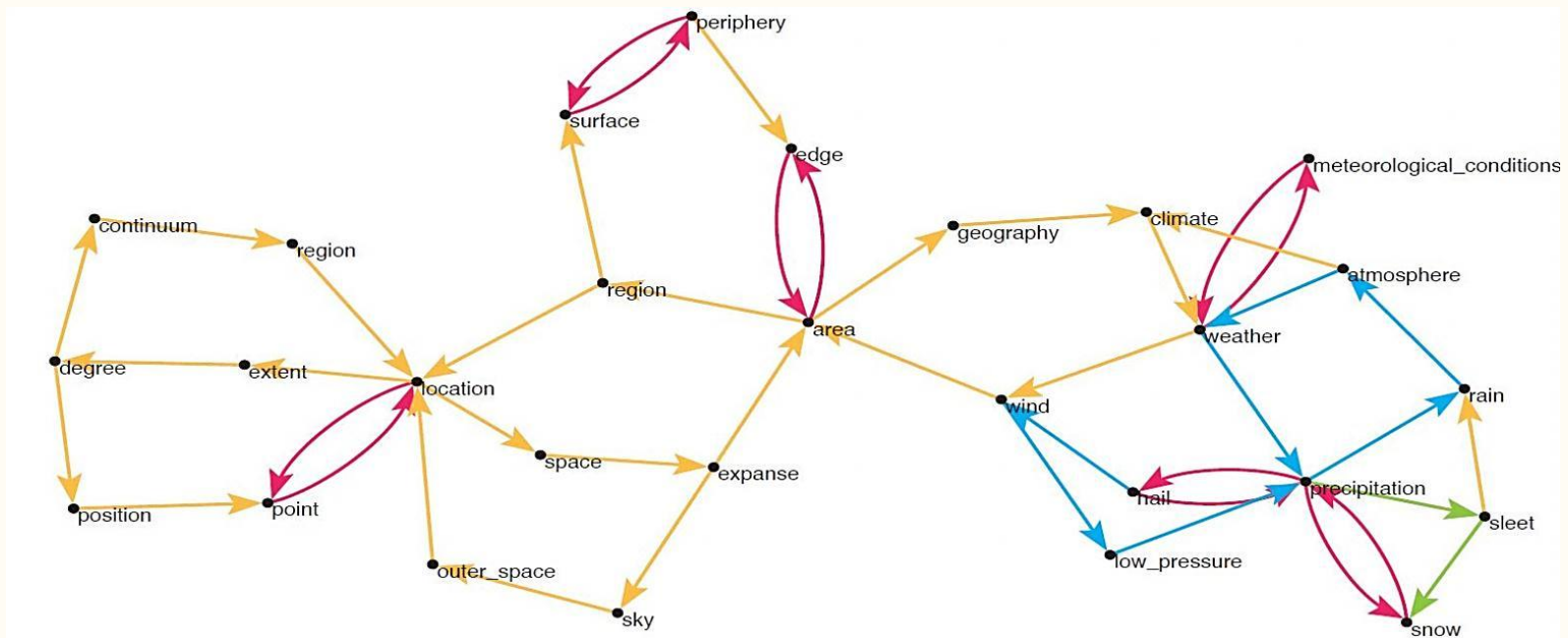
Loops and self reference in the construction of dictionaries

Rutgers 2012

J-P. Eckmann, D. Levary, E. Moses & TT

Outline: Dictionaries as loopy graphs

- Dictionaries as graphs.
- “Loopiness” of dictionaries.
- Motivation: self-reference, loops, emergence...
- Possible link to “concepts” and meaning.
- Structure of dictionaries (core, loops clusters).
- Evolution of dictionaries (historical).

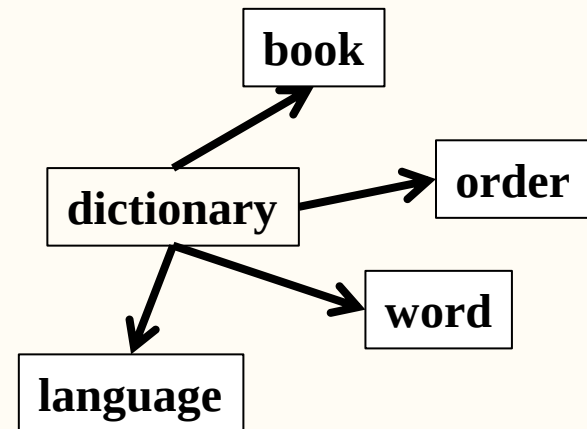


Graphs are natural representation of dictionaries

- New words added to language\ dictionaries to:
 - (i) increase efficiency of communication
 - (ii) introduce new concepts...
- Binary relations (e.g. synonyms, general-specific) $\rightarrow$ edges in word graphs.
- **Dictionary graphs** $G(V, E)$ – arrows point from word to words in its definition:
 $V = \text{defined words}; \{v_1 \rightarrow v_2\} \in E \text{ iff } v_2 \in D(v_1).$

Dictionary (n) A *book* which explains or translates, usually in alphabetical *order*, the *words* of a *language*.

$D(\text{dictionary}) = \{\text{book, order, word, language}\}$

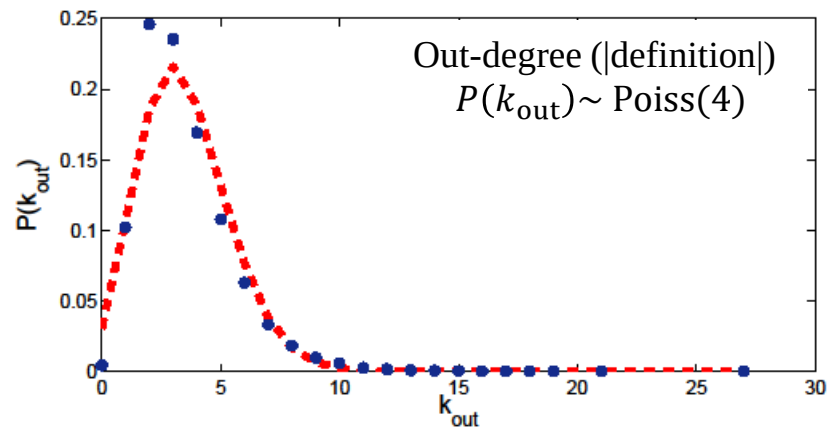
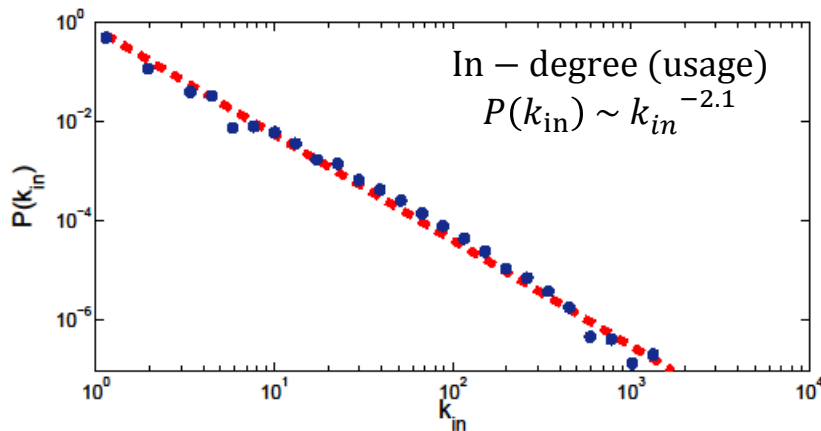


Dictionary graphs are highly connected

- Main example: (eXtended) **WordNet**®
- Synsets (grouped synonyms, reduces noise).
- 80,000 synsets (of nouns).
- 300,000 directed edges.
- Results in a large directed graph.



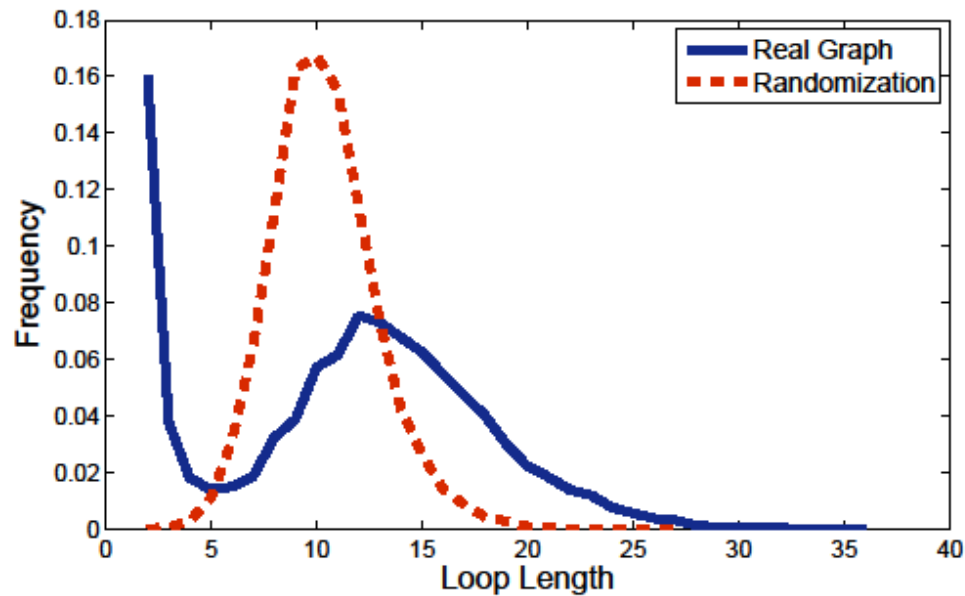
(Fellbaum et al.)



(Dictionary graphs discussed by Litkowski (1970s) and recently by Harnad et al.)

Dictionary graphs contain many short loops

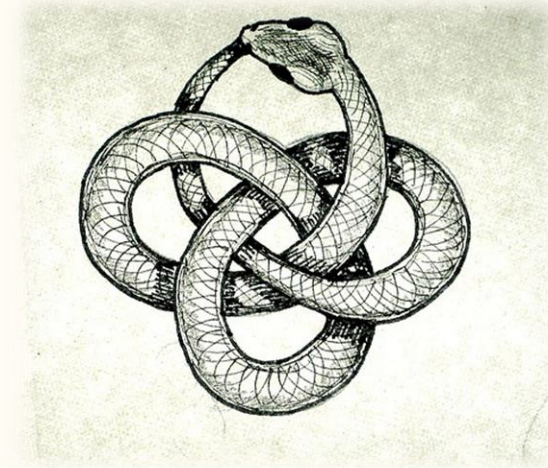
- Compared to randomized graph with same degrees.



?

Theoretical motivation: why dictionaries are loopy?

- **Loops**: from classical paradoxes to self-replication in biology...
related to emergence...
- **Dictionaries** are inherently cyclic (all words are defined),
much more than randomized graphs.
- Possible roles of loops in structure and evolution of dictionaries (languages)?
- Suggestion: **Loops are not artifacts but related to emergence of *new concepts*.**
- Philosophical problem: What is a concept?
(e.g., Kant: concept results from mental operations on set of “images”.
→ Concept emerges when minimal set of “images” accumulates.)



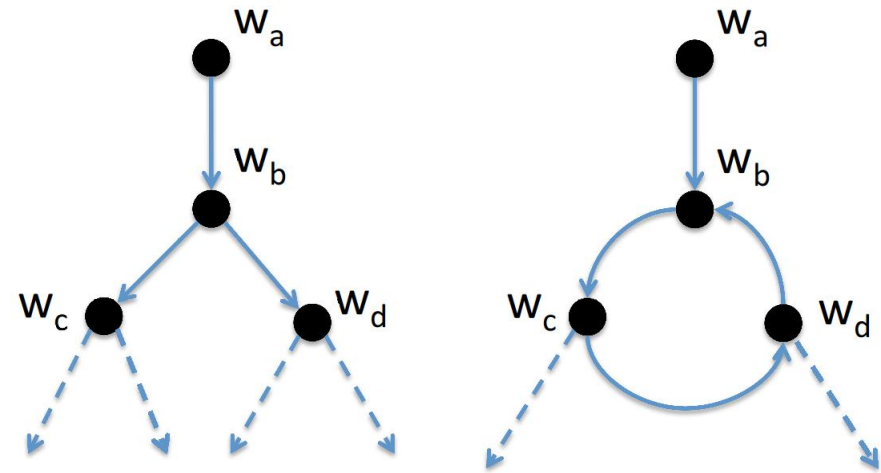
Is emergence of loops also emergence of concepts?

- Time series of dictionary construction $W = \cup_t W_t$.
- Emergence of new “concept” at $t_0 \equiv$ word w was *undefinable* at $t < t_0$.

- $D(w) = \{\text{words in definition of } w\}$.
- $D(D(D(D\dots)))) = \text{walking along definitions}$.

- Word can be replaced by definition:

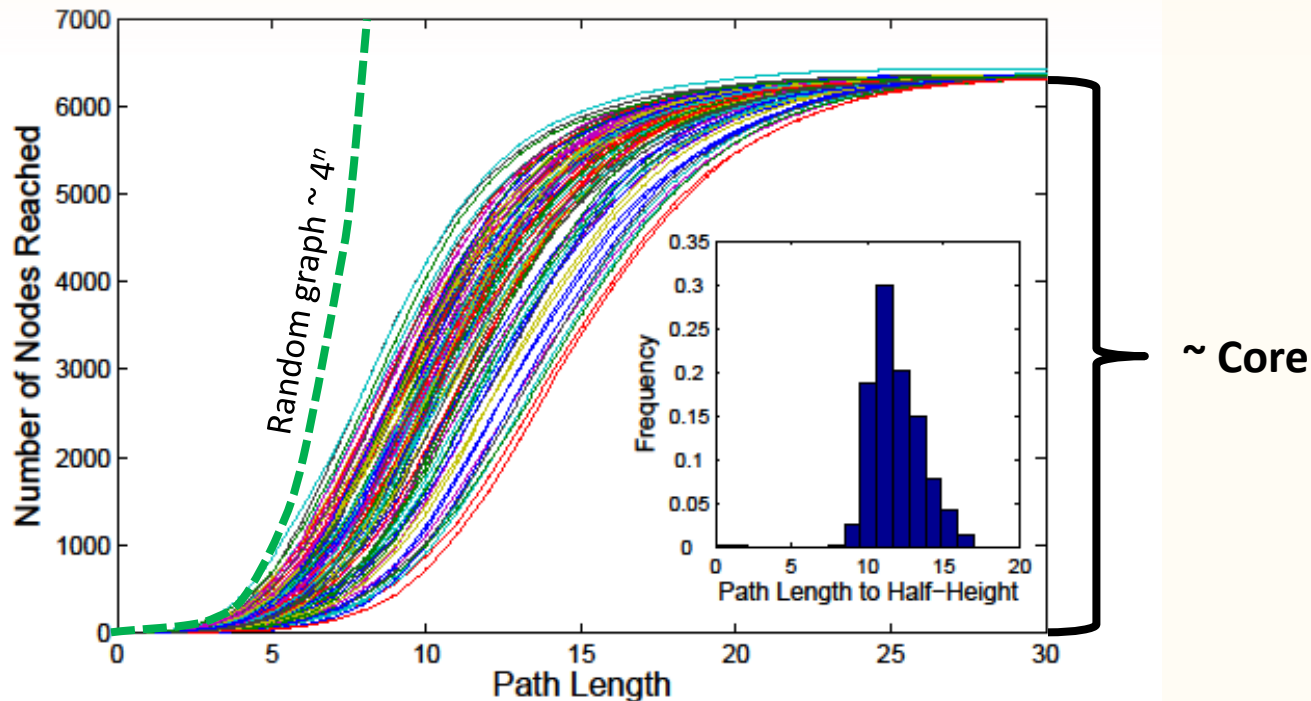
$$w \rightarrow D(w) \rightarrow \dots \rightarrow D^{(n)}(w).$$



- G is tree: translate to older words, everything is definable.
- Loop cannot be reduced to older words: undefinable words $\rightarrow$ new concepts.

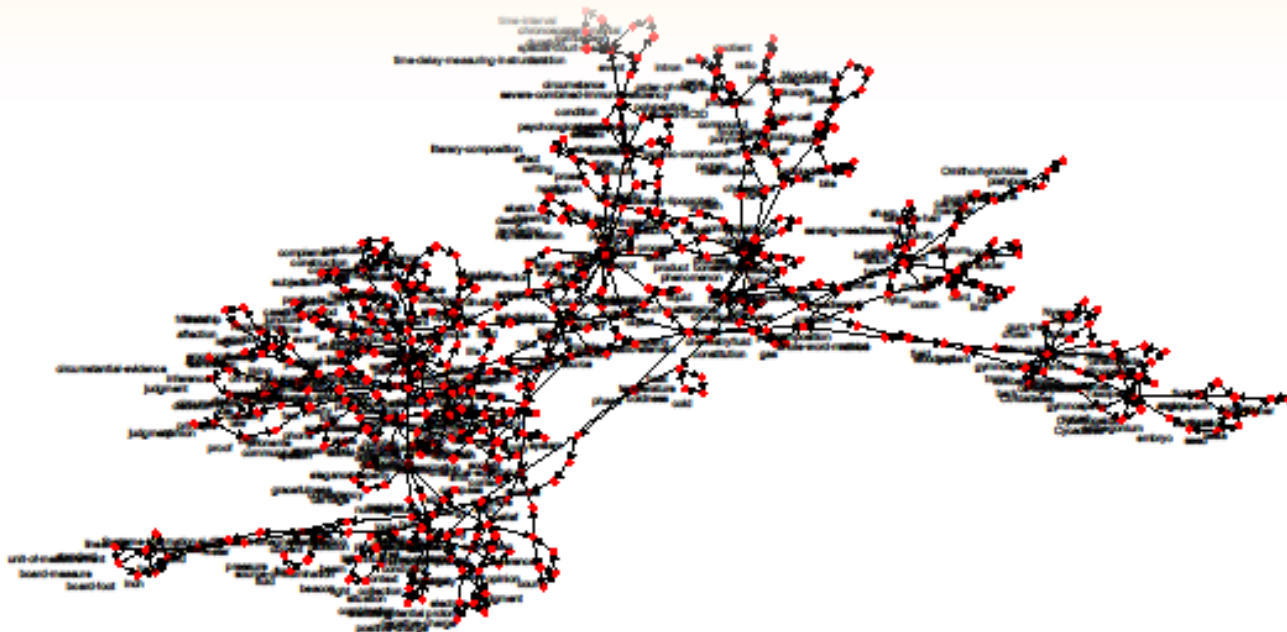
The dictionary converges into a “core”

- eXtended WordNet resolves polysemous words (same meaning).
- Examine only nouns (most related to concepts).
- Definition paths converge *fast* into “**core**” =
~ 6,300 words reachable from 99% of words.
- Graph is (i) *not tree-like* (ii) *not small-world* (large diameter).



The core is made of loops

- **Core** is obtained by “*shaving hairs*” (words not used in definitions).
- Shaving seven times leaves only loops.
- ~9000 words are in loops.
- ~1,100 strongly connected component (everybody reaches everybody).
- Most components are small but one has 6300 nodes = the “core”.



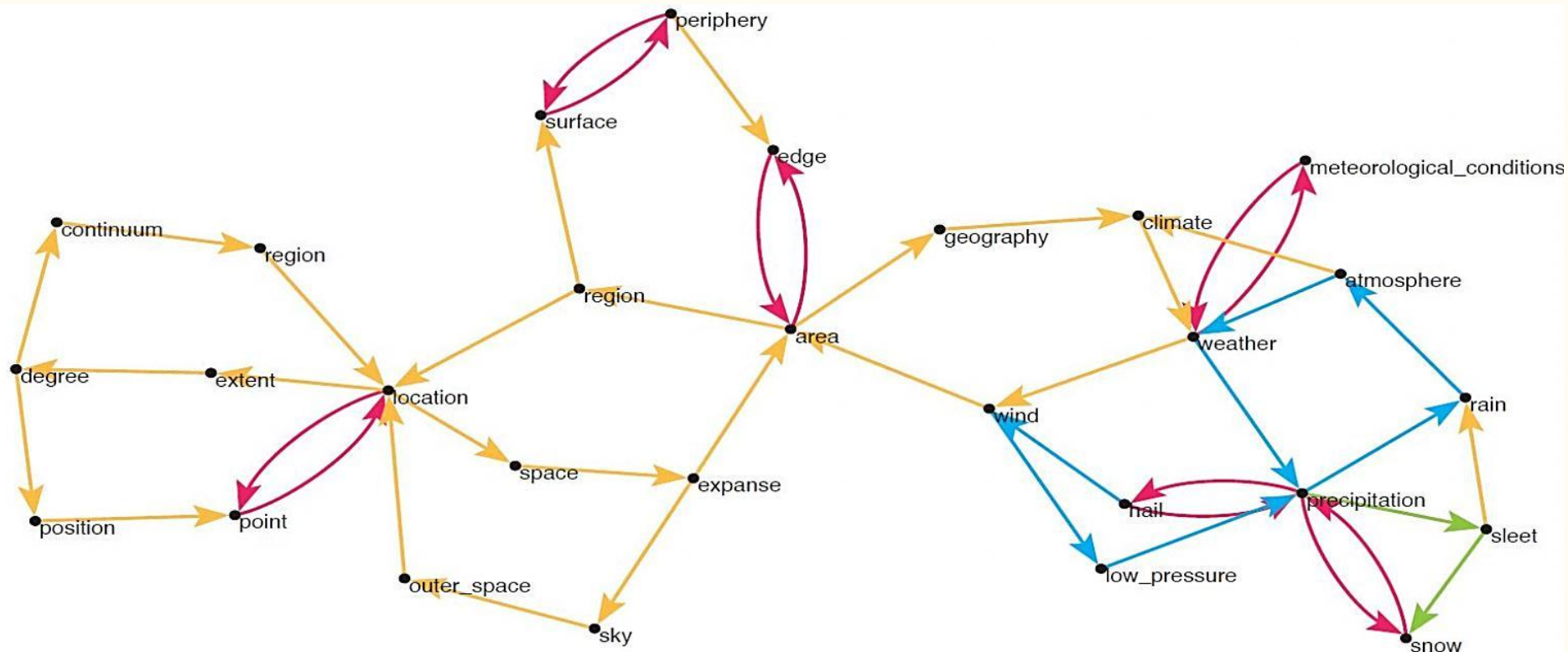
Cores are quite universal and overlap with most useful words

	Core	Ogden (600)	Jōyō Kanji	Gutenberg
Core	1595	314 (52%)	403 (29%)	265 (39%)
Basic English		600	328 (24%)	213 (32%)
Jōyō Kanji			1376	319 (47%)
Gutenberg				673

- Ogden’s “**Basic English**”: 850 words for daily discourse.
- **Joyo Kanji**: the Japanese Education Ministry's list of essential 1,376 characters.
- **Gutenberg**: top 1000 words in all books on Project Gutenberg.
- Missing words are “over-specific” yet useful words: apple, brick, chalk...
 $\text{apple} \in \text{Ogden}, \text{ but } D(\text{apple}) = \{\text{red, fruit}\} \in \text{Core}.$

The loop structure of the core

- Focus on **core** (largest connected component) =
Directed graph where every node is in a loop and any two nodes are connected.
- Assign to nodes\edges size of shortest loop.

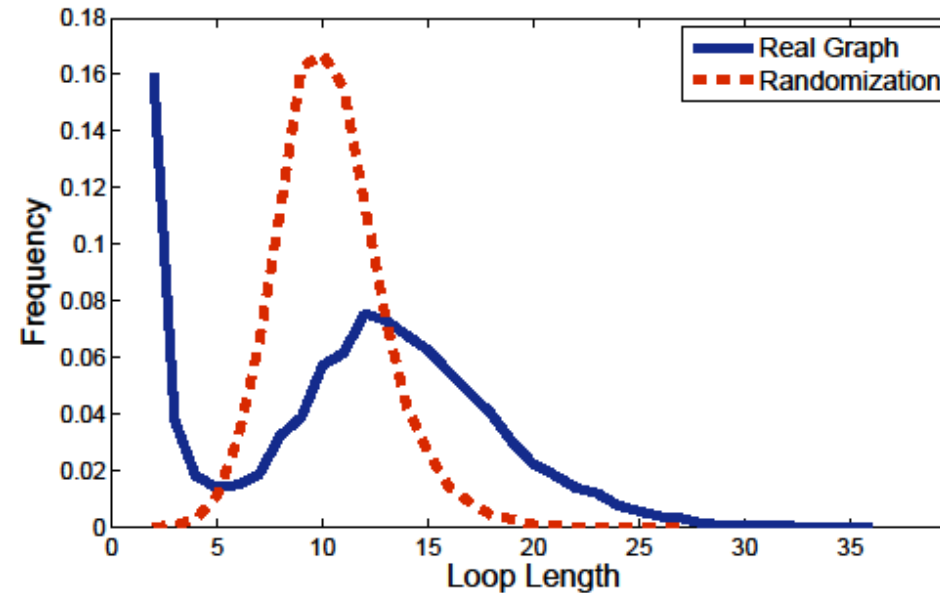


- Large connected component Red=2, Green=3, Blue=4, Yellow=5.

The loop structure of the core

- Long loops are susceptible to errors:

railcar → rails → bar → weapon →
→ instrument → skill → train → railcar.

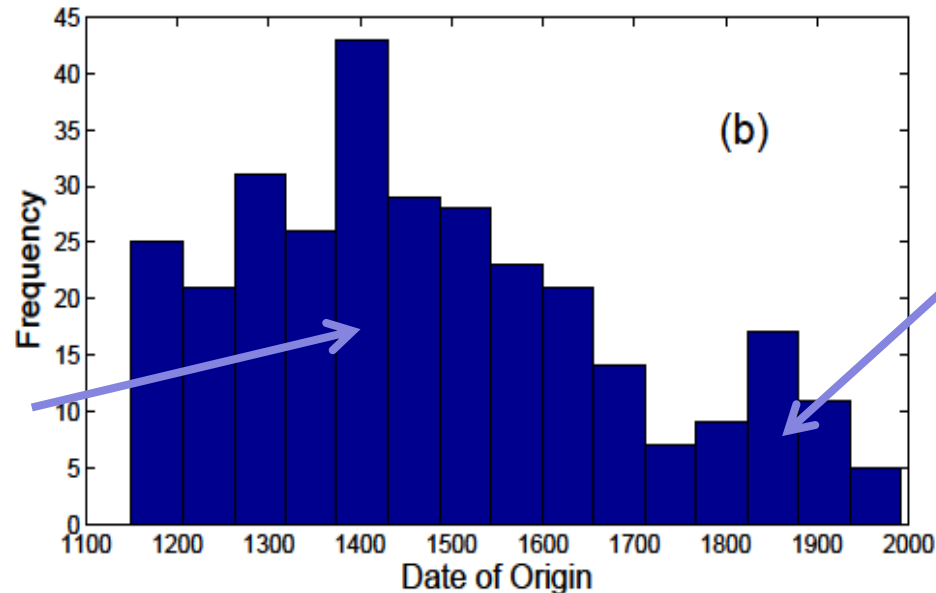


emotion	height	bark	injury	winner
spirit	end	trunk	violence	contestant
dejection	dimension	tree	accident	competition
melancholy	length	lumber		
feeling				

- Examples of strongly connected components.

Dynamics of dictionary construction

- Date of word's first appearance (online etymology dictionary).
- Approximated order of construction $W = \cup_t W_t$.
- Temporal loop distribution exhibits two distinct peaks:

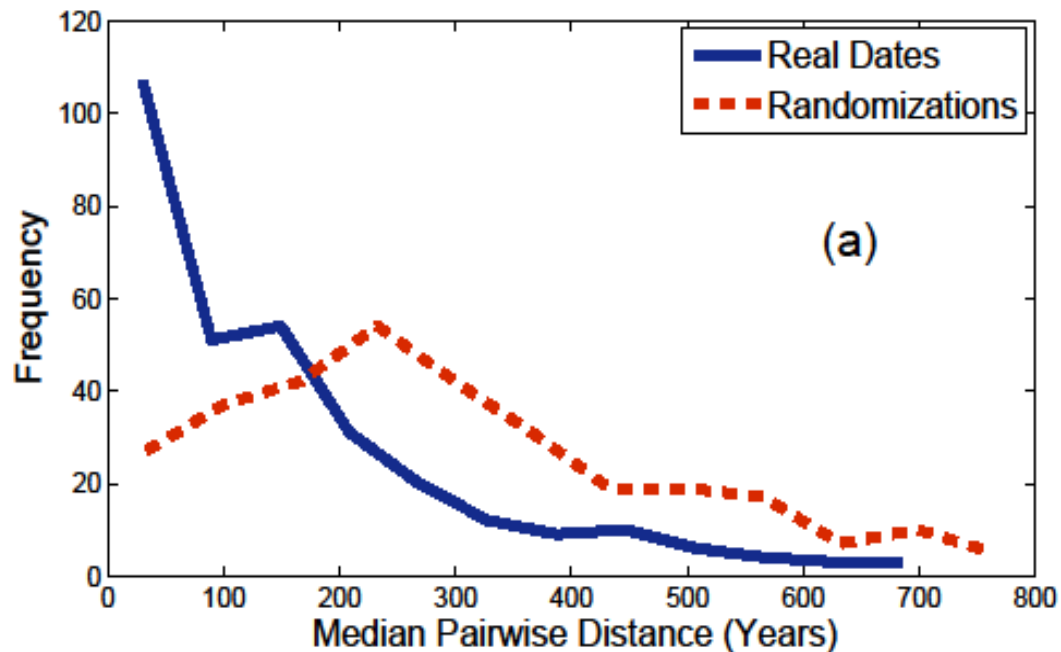


14th-16th century:
Middle → Modern English

19th-20th century:
science & technology

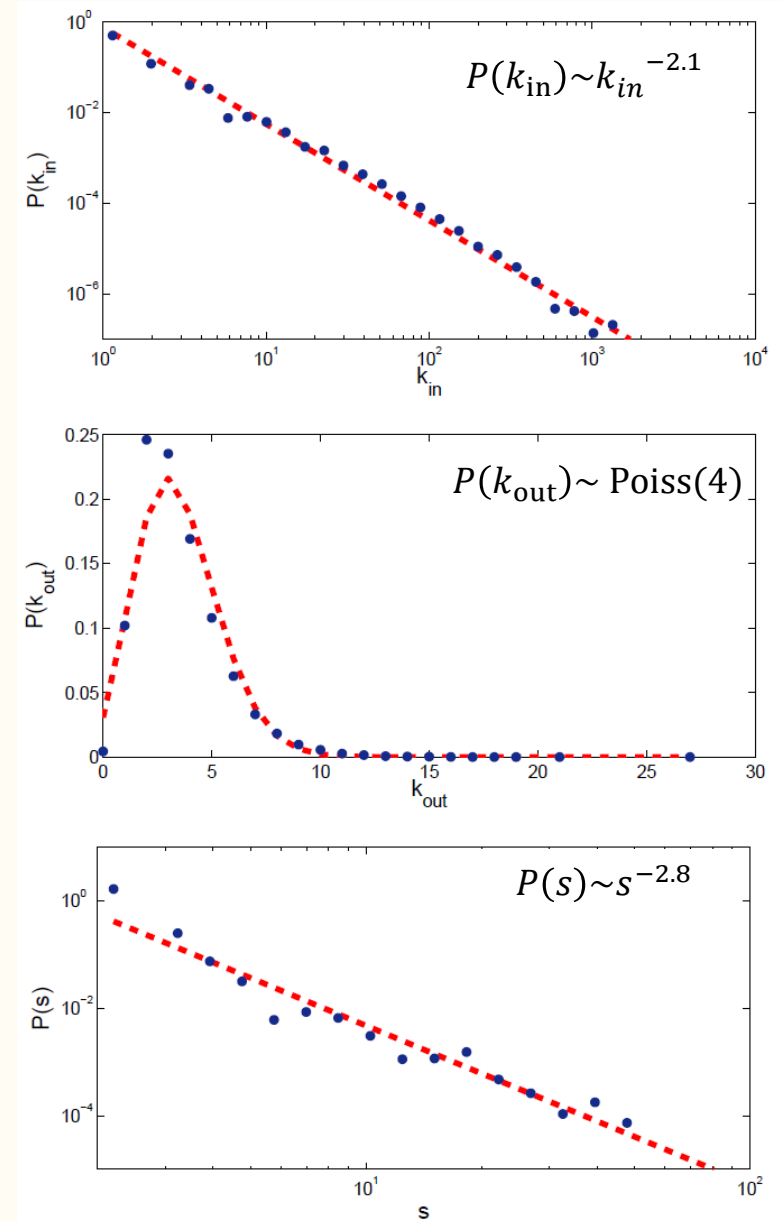
Loops tend to appear “simultaneously”

- Median pairwise distance between words in a loop.



Scale-free loop distribution

- Distribution of loop size $P(s) \sim s^{-2.8}$.
- Can be “explained” by preferential attachment:
Words and loops attach in proportion to size.



Concept “basis”? Large-scale structure?

- Loops are linked to concepts.
- What is the large-scale structure of concepts?
- “Decomposition” of words:
 proximity to each of ~400 short (<5) loops (measured by # of short paths).
- Proximity matrix 80000 x 400 → SVD → reveals “eigenmodes” of loop clusters.

Vector 1	Vector 2	Vector 5	Vector 6	Vector 8
<i>Old World, oceans</i> <i>bodies of water</i> <i>The Americas</i> <i>influence, power</i>	<i>spine, brain</i> <i>body parts</i> Old World, oceans <i>energy</i> pathology <i>narrative</i> <i>cognition</i> <i>organic process</i> <i>respiration</i>	grains <i>bodies of water</i> herbage flower, seed <i>land</i> <i>water</i>	<i>body parts</i> <i>flower, seed</i> <i>tree, bark</i> grains nucleus, DNA spine, brain <i>Gymnosperms</i> bodies of water pathology	Jesus, Christianity man, woman <i>nucleus, DNA</i> Roman Empire student, teacher speaker, speech Vatican, absolution Old Testament book

Table 3: Examples of the highest singular components for the dictionary. The elements in the singular components are the semantically cohesive clusters of words obtained from decomposing the core. For table entries, word(s) representing the main theme of each cluster were chosen. Clusters are listed in order of the absolute value of their coefficient in the singular component. Only components whose absolute values were greater than 0.1 were listed. Plain text and italics indicate positive and negative values respectively. The 3rd, 4th, and 7th highest singular components were very similar to the vectors shown and therefore not displayed.

Summary

- Loops appear as simple semantic units of dictionary graphs.
- Make a highly connected “core” of basic words.
- Relation to meaning: New loops $\leftrightarrow$ New “concepts”.
- Tend to appear “simultaneously” (following historical dictionaries)
- SVD: eigenmodes of loops (concepts)?
- Overall structure of loops reflects semantic features, esp. if loops are labeled by another “coordinate” like "time"
-

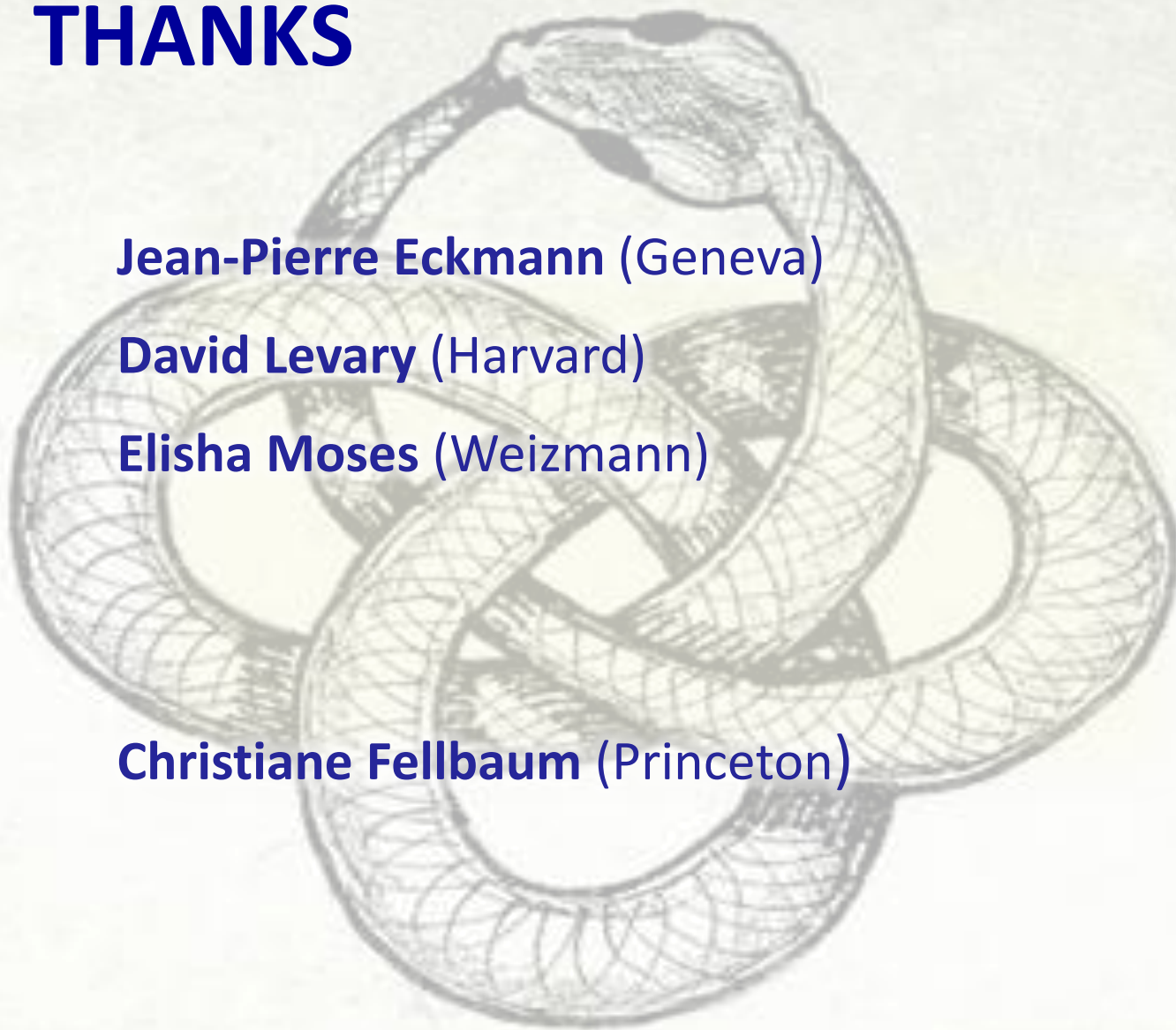
THANKS

Jean-Pierre Eckmann (Geneva)

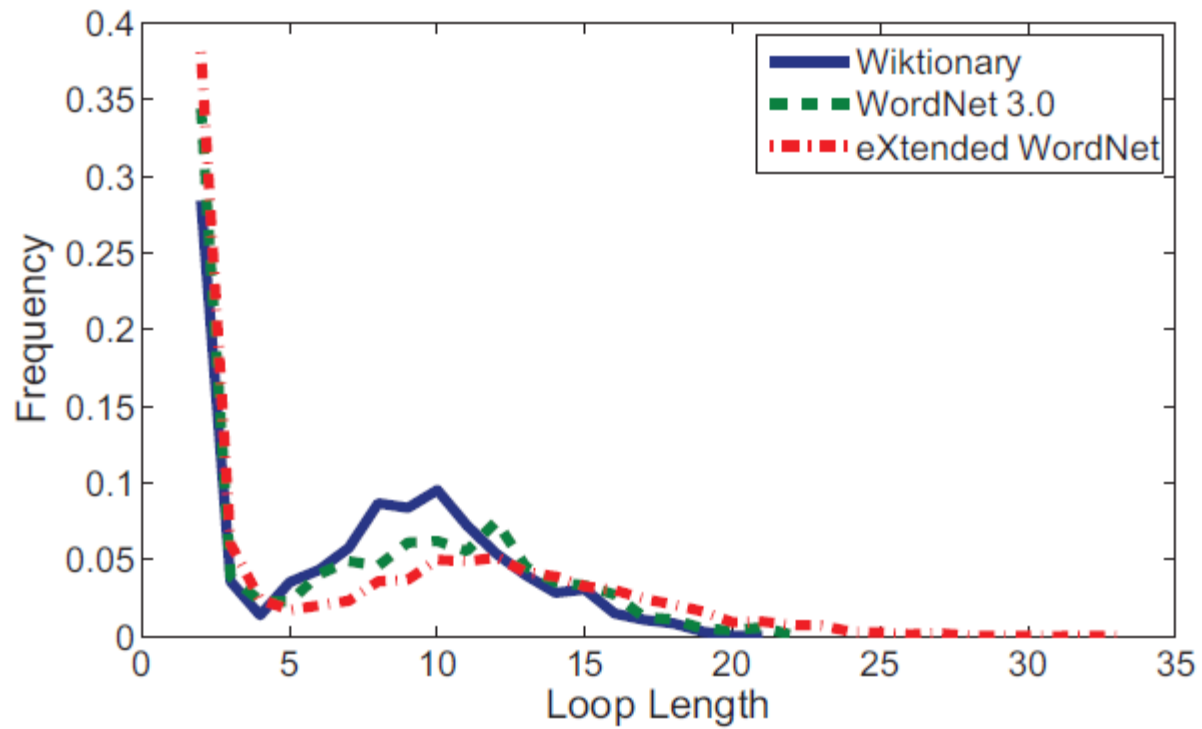
David Levary (Harvard)

Elisha Moses (Weizmann)

Christiane Fellbaum (Princeton)



Tests



WordNet tree vs. Loops

Wheel

=> machine, simple machine

=> mechanical device

=> mechanism

=> device

=> instrumentation

=> artifact

=> whole, unit

=> object

=> physical entity

=> entity

2 axle->wheel->axle

6 car->wheel->vehicle->transport->transportation->passenger->car

6 wheel->vehicle->transport->transportation->passenger->car->wheel

7 wheel->can->container->transport->transportation->passenger->car->wheel

9 conduit->tunnel->car->wheel->disc->needle->leaf->transpiration->tube->conduit

9 disc->needle->leaf->transpiration->tube->conduit->tunnel->car->wheel->disc

9 leaf->transpiration->tube->conduit->tunnel->car->wheel->disc->needle->leaf

9 needle->leaf->transpiration->tube->conduit->tunnel->car->wheel->disc->needle

9 shaft->form->can->container->transport->transportation->passenger->car->wheel->shaft

9 transpiration->tube->conduit->tunnel->car->wheel->disc->needle->leaf->transpiration

9 tunnel->car->wheel->disc->needle->leaf->transpiration->tube->conduit->tunnel

9 wheel->disc->needle->leaf->transpiration->tube->conduit->tunnel->car->wheel

9 wheel->shaft->form->can->container->transport->transportation->passenger->car->wheel