Semantic Networks Generated from Early Linguistic Input

Andrei Amatuni and Elika Bergelson | Department of Psychology and Neuroscience, Duke University

Research Questions:

1) Do common semantic network properties *necessarily* stem from **incremental growth**? (preferential attachment vs. preferential acquisition)

2) Does a word's node degree *correlate with age of acquisition* in networks built using a **static metric of semantic similarity (GloVe)**

Background

Structure in Semantic Networks

- Common structure observed across different semantic nets
  - **scale free** degree distributions [**P(k) ~ k$^{-\alpha}$**]
  - **small-world** organization [**L ∝ log *N***]
  - **high clustering** coefficients

- **Incremental network growth** proposed as the cause of **scale free** network structuring *in general* (Barabási & Albert, 1999),

common examples:
- world wide web
- social networks
- citation patterns in scientific publications

This incremental model uses **preferential attachment:** *new nodes are more likely to get added to more connected nodes*

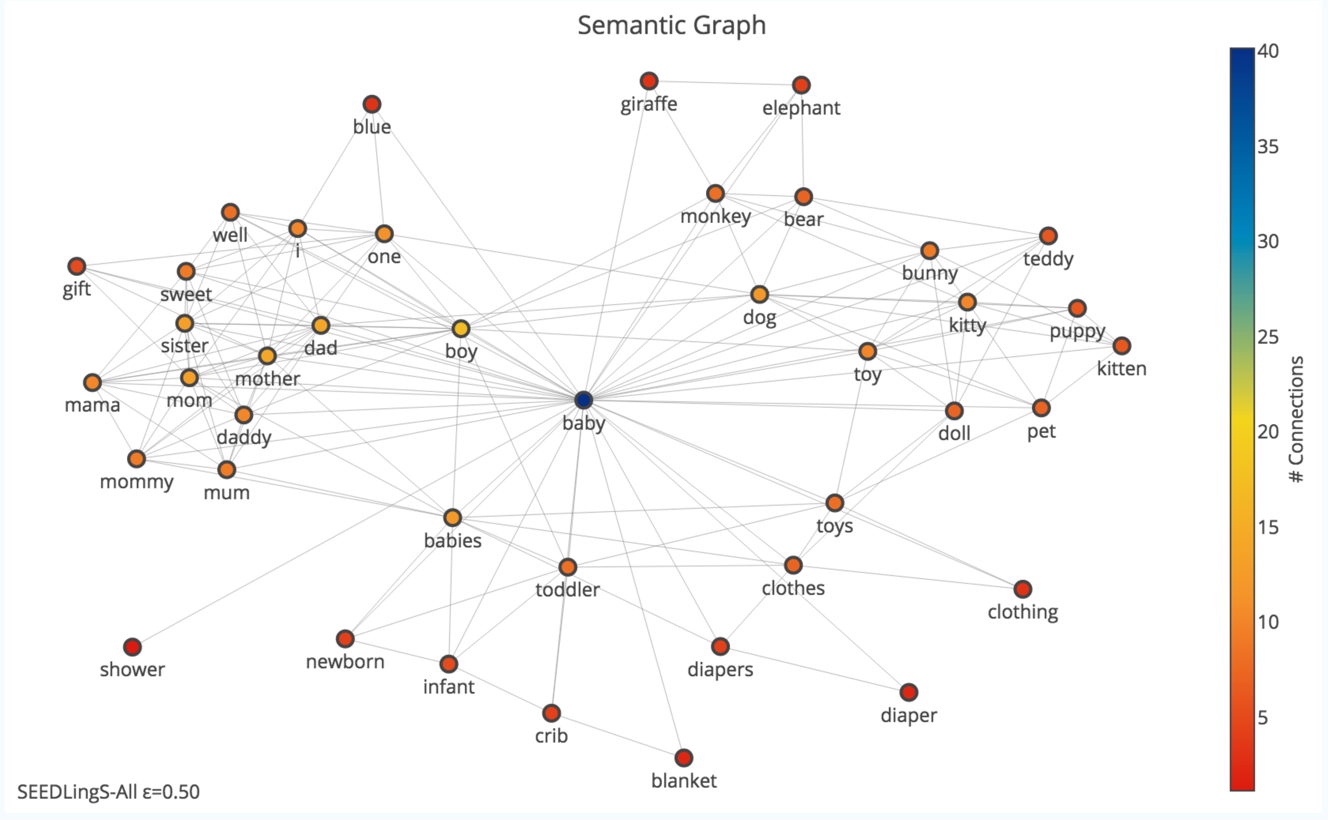
- Incremental growth assumed to correspond to age of acquisition for words.

- Steyvers and Tenenbaum (2005): compared ***incremental*** networks vs. ***all at once*** networks (LSA, i.e. semantic vector space model)
  - LSA Networks lacked common semantic net features
    - taken as support for incremental growth *leading* to common net features

- **Incremental** models assume semantic similarity is ***relative in time***
  - newly learned word has different semantic neighbors as a function of the state of the lexicon during learning

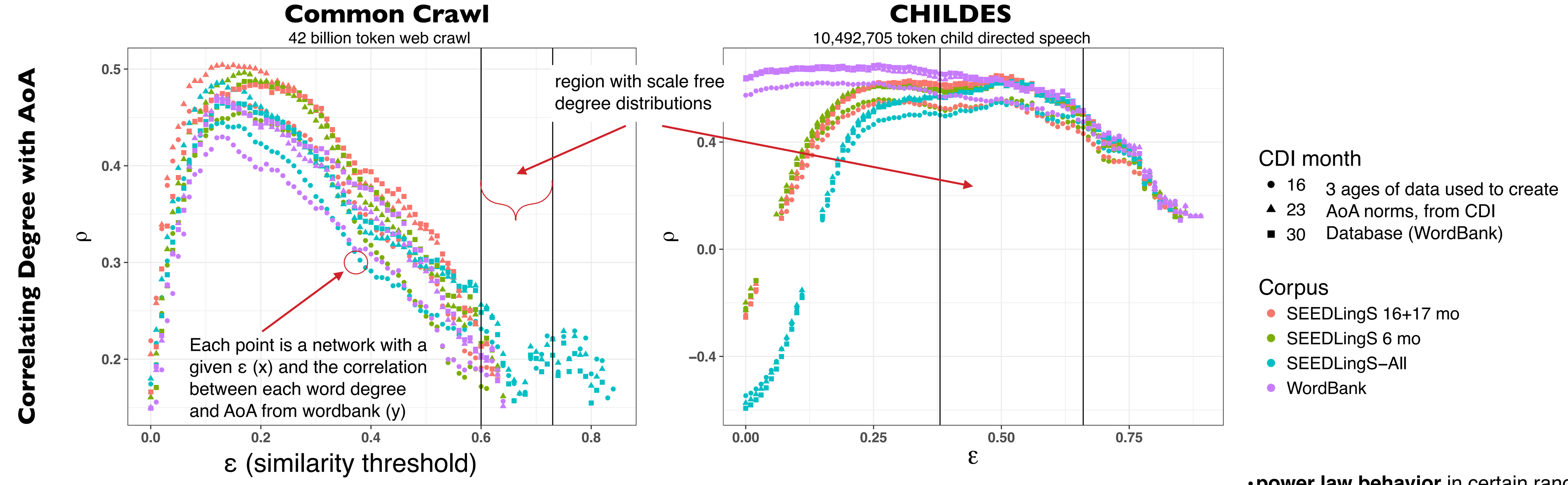
- Hills et al. (2009): counterproposal: **preferential acquisition.**
  - semantic structure in *environment* guides acquisition, not structure in existing lexicon
  - i.e. the 'ground' of semantic similarity is independent of the learner

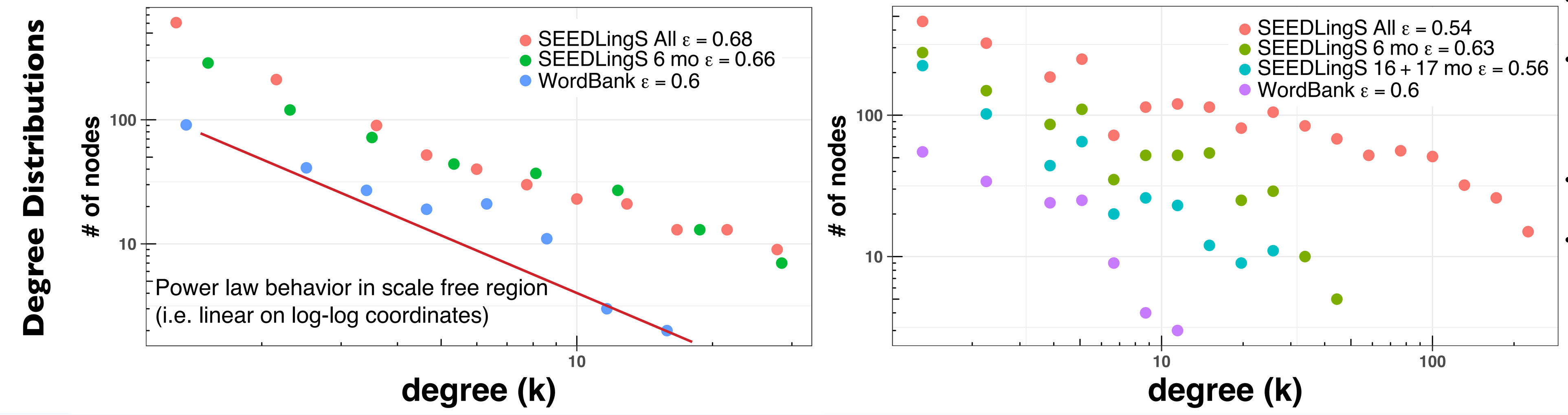
We use new generation **all-at-once** (i.e. non-incremental) networks (GLoVe), and a large new corpus of nouns heard by infants (**SEEDLingS**) to test limits of previous claims.



Sample Semantic Net for "baby"

Results





- **power law behavior** in certain ranges of ε
- **Node degree correlates with AoA** across:
  - wide range of ε
  - similar for each CDI month
- Cross corpus behavior very similar in C.C.
- advantage for WordBank in CHILDES training network

Clustering Coefficients and Average Shortest Path

	Corpus	ε (at peak ρ)	Clustering Coefficient	Avg. Path Length (L)
Common Crawl	SEEDLingS All	0.13	0.594	1.749
	SEEDLingS 6 mo	0.16	0.669	1.739
	SEEDLingS 16+17 mo	0.12	0.726	1.534
	WordBank	0.13	0.895	1.202
	Erdős–Rényi	-	**0.049** **random network baseline**	**1.950**
	Watts-Strogratz	-	**0.634** **prototypical small-world**	**3.013**
CHILDES	SEEDLingS All	0.52	0.273	4.971
	SEEDLingS 6 mo	0.52	0.264	5.614
	SEEDLingS 16+17 mo	0.49	0.266	4.961
	WordBank	0.26	0.479	1.866

Methods

Defining Semantic Similarity

- Common approaches:
  - Thesaurus
  - WordNet
  - Adult free association norms

- Common approaches all lead to scale-free distributions, small-world structure, clustering
- All these methods use ***static metrics of semantic similarity***

We use **GloVe** (Pennington, Socher, & Manning, 2014), a new semantic vector space model, as our similarity metric.

- word vectors based on ratio of word co-occurrence probabilities for a given training corpus
  - Results in a ***static geometric encoding*** of semantic similarity

New Approach: similarity with GloVe:
cosine between 2 word vectors, relative orientation in high dimension vector space (d = 300)
- cos(θ) = 1: identical words
- cos(θ) = 0: orthogonal words
- cos(θ) = -1: words pointing in exactly opposite directions

Networks built with **SEEDLingS:** a newly collected corpus of early linguistic input to children.
- All constituent nodes in our graphs are ***nouns infants hear and attend to in their daily lives***.
- 3 subsets of corpus:
  - **6 month (**1855 unique noun types, 29289 unique noun tokens)
  - **16+17 month** (1708 unique noun types, 26969 unique noun tokens)
  - **All months** (4359 unique noun types, 194204 unique noun tokens)
  - **Wordbank** (369 unique words) as comparison network

- Each word has a corresponding GloVe word vector
- Iterate through all words:
  - draw edges (neighbors) if cosine between them is within ***threshold (ε)***
  - incrementing ε by 0.01, resulting in 100 different networks for each corpus subset

Discussion

1) Do common semantic networks properties *necessarily* stem from **incremental growth**?

- **No**, using GLoVe vectors to build semantic networks using a **static** metric of semantic similarity (i.e. non-incremental nets), we find:
  - ***scale-free***, ***small-world***, and ***highly clustered*** semantic networks
    - evidence against strong 'incremental' claim of preferential attachment

2) Does a word's node degree correlate with age of acquisition in networks built using a **static metric of semantic similarity (GloVe)**

- **Yes**, depending on ε (similarity theshold), medium correlations between node degree & AoA (Spearman's $\rho \sim 0.5$, $p < 0.05$)
- Node degree also correlates with **frequency** in corpus (see paper for details)
  - **frequency and node degree together** accounts for significantly more variance than either alone in predicting word production

Conclusions
- We can build scale-free semantic networks using an **all-at-once** method, defining similarity in terms of a geometric encoding of distributional information.
- The original failure to do so using LSA is not indicative of hard constraints on the mechanisms responsible for structuring.
- Compatible with 'inherent' semantic structure in the external world (à la Hills et al, 2009).
- infants and caregivers may be sensitive to this nonuniform distribution of semantic information (future work needed).

Ongoing & future directions
- analyze difference between algorithms trained on large internet corpora (Common Crawl) and those that are child specific (CHILDES)
- explore random walks on semantic networks
- model the external semantic networks as a generative process

References

- Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.
- Bergelson, E. (2016a). Bergelson seedlings homebank corpus. doi: 10.21415/T5PK6D
- Bergelson, E. (2016b). Seedlings corpus. Retrieved 2017-01-29, from https://nyu.databrary.org/volume/228
- Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman,V. A. (2016). Wordbank: An open repository for developmental vocabulary data. Journal of child language .
- Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic net- works preferential attachment or preferential acquisition? *Psychological Science*, *20*(6), 729–739.
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical methods in natural language processing (emnlp)* (pp. 1532–1543).
- Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, *29*(1), 41– 78.