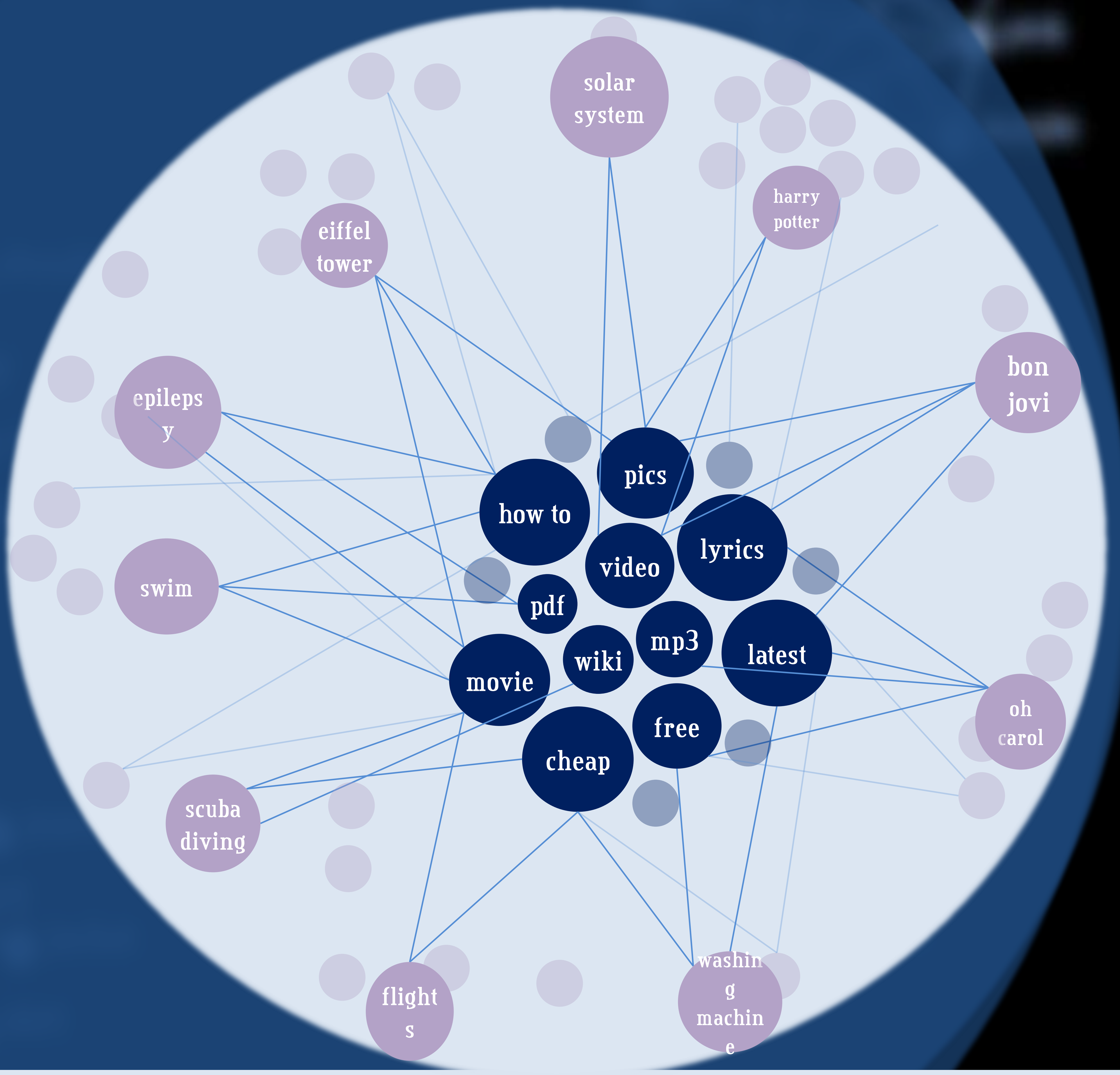
Complex Network Analysis of Word Co-occurrence Networks

Natural Language	Queries
|K| = 5000 units	|K| = 1000 units
|P| = 84,000 units	|P| = 1,200,000 units
Sentences formed by units from K and P, or only K	Queries mostly formed by units from K and P, or only P
Intra-K edges dominate	K to P edges dominate
K more tightly coupled	K less tightly coupled

Property	NL	Queries
Degree Distribution	2-regime	2-regime
Clustering coefficient	0.437	0.630
Average shortest path	2.670	3.305



Structural Complexity of Web Search Queries

through the Lenses of Positionality, Language Models and Networks

Rishiraj Saha Roy and Niloy Ganguly, (IIT Kharagpur, India) and Monojit Choudhury (Microsoft Research India)

Change in Segment Positions

- For 2006 and 2010 logs, segments with the highest co-occurrence counts are labeled *intent*, and the rest as *content*
- For each segment, query beginning probability P_b, ending probability P_e and that of occurring in the middle P_m are computed
- Navigational queries like *imdb* and *youtube* are now appended as intent (P_b drops)
- Intent segments (*how to, news*) stabilizing towards ends of query (P_b rises or P_e rises)
- Stacking of intent segments gradually making search queries longer

Perplexity of Language Models

Model	NL	Queries	NL	Queries
	(Perplexity)	(Perplexity)	(Counts)	(Counts)
1-gram	1,406.59	6,417.28	0.3M	0.2M
2-gram	193.722	104.337	3.5M	1M
3-gram	17.663	5.43	9.7M	1.1M
2-set	893.851	384.945	48.1M	4.2M
3-set	N.A.	23.36	N.A.	24.8M

- Perplexity is an information theoretic measure of how perplexed a user is in predicting the n^{th} word
- Perplexity of unigram model much larger for queries
- In contrast, bigram and trigram perplexity much lower for queries`

Conclusions

- Web search queries provide a very interesting case of a complex self-organizing communication system which has its unique characteristics
- Queries structurally simpler than NL, but more complex than bags-of-words model
- Several similarities with NL that make this system interesting to study from a language evolution perspective

Contacts: rishiraj.saharoy@gmail.com
niloy@cse.iitkgp.ernet.in
monojitc@microsoft.com