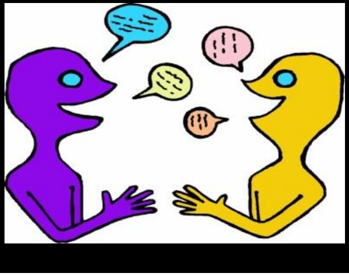
Complex Network Analysis Reveals Kernel-Periphery Structure in Web Search Queries

Rishiraj Saha Roy and Niloy Ganguly
IIT Kharagpur
India

Monojit Choudhury
Microsoft Research India
India

Naveen Kumar Singh
NIT Durgapur
India

Query Representation and Understanding Workshop 2011 (QRU '11)
ACM SIGIR 2011, Beijing, China



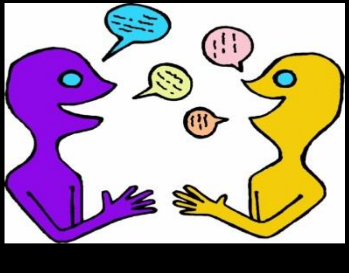
Language of Queries

- Interaction between user and search engines over the years has resulted in the evolution of a distinct language for Web search queries

✓ *gprs config samsung focus at&t*

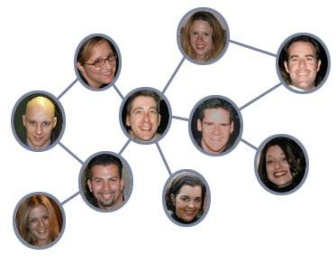
✓ *samsung focus at&t gprs config*

✗ *focus config at&t gprs samsung*



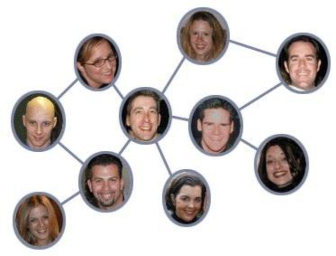
Language of Queries

How can we begin to analyze this new language?



Complex Networks

- Real life networks not easily explained by standard topologies
- Applications to linguistics – word co-occurrences, consonant inventories, syntactic and semantic features, language dynamics



Complex Networks

Word co-occurrence networks: Interesting tool to discover fundamental properties of a language

bing™

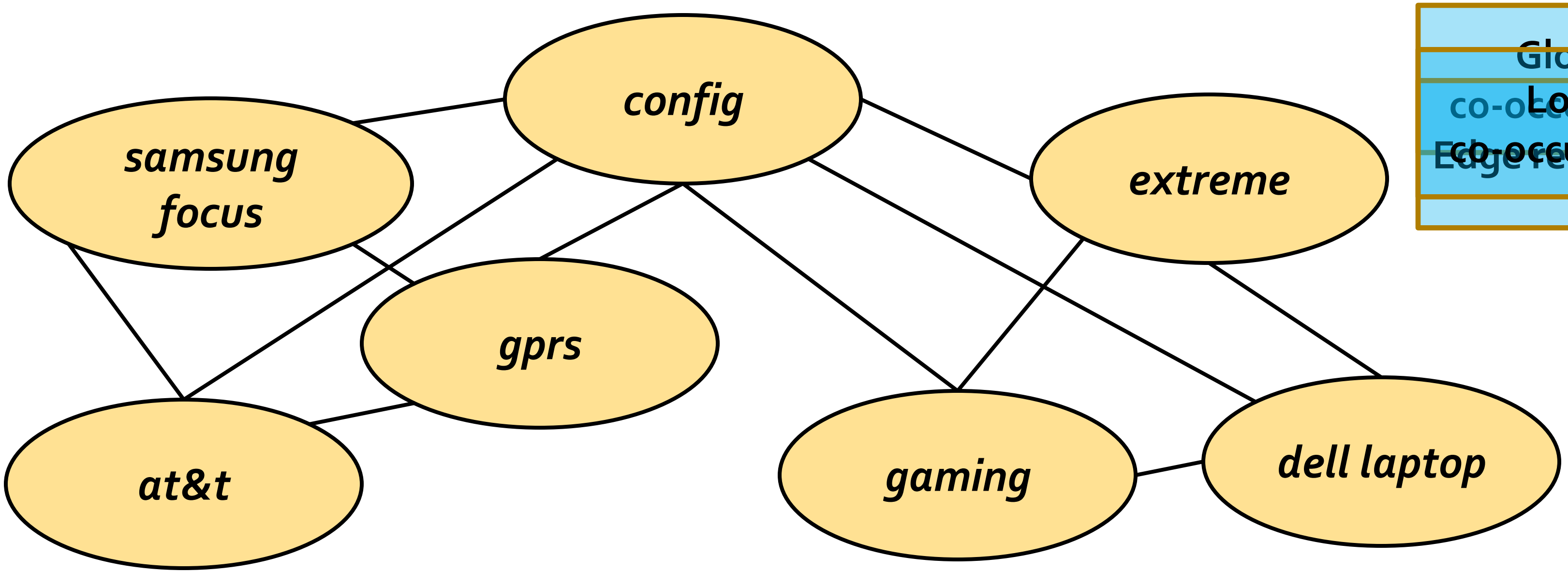
Data

16.7 million entries sampled from Bing Query Logs from Australia (February – May 2009)

Courtesy: Microsoft India Development Center

Network Models for Queries

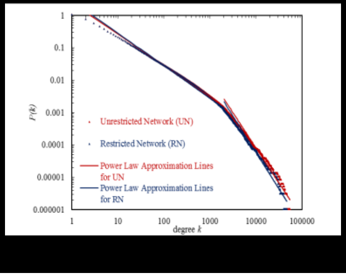
- *"gprs" "config" "samsung focus" "at&t"*
- *"dell laptop" "extreme" "gaming" "config"*



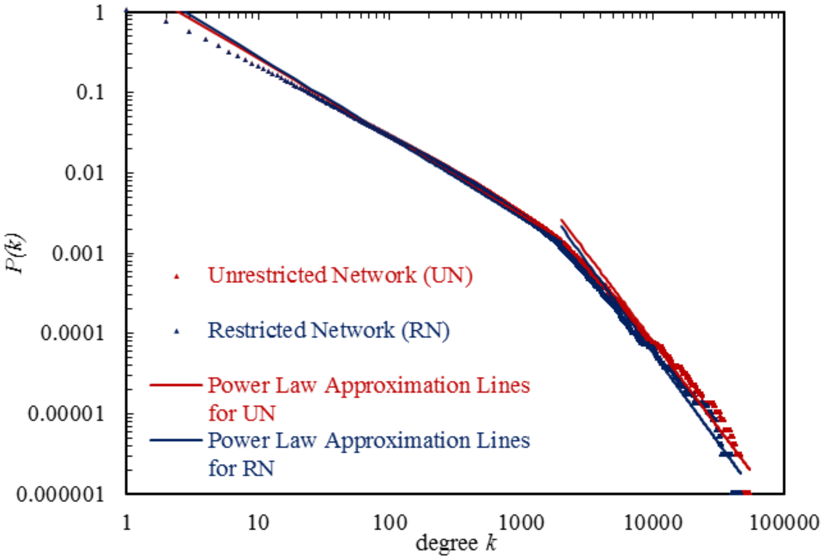
Global co-occurrence

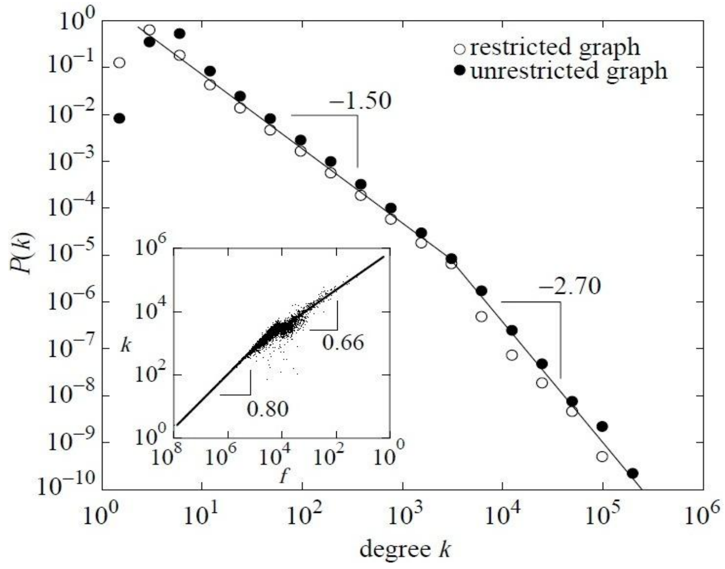
Local co-occurrence

Edge restriction

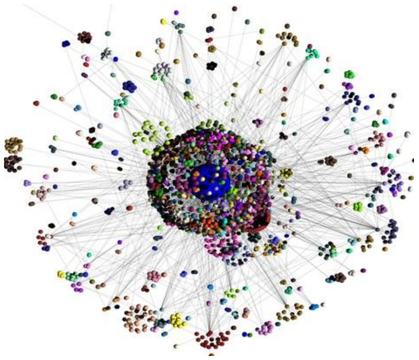


Two-regime Power Law





- Two-regime power law in degree distribution
- Similar coefficients for queries and English
- Kernel (K-Lex) and peripheral (P-Lex) lexicon distinction

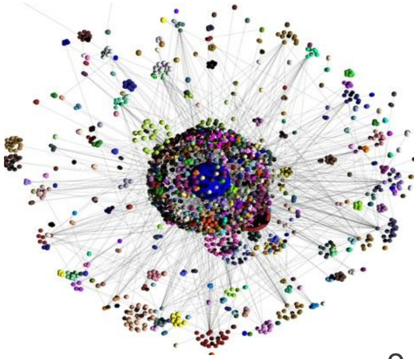


Insights (1)

- ✓ K-Lex and P-Lex
- ✓ Higher mean shortest paths
- ✓ Less tight kernel
- ✓ More k-p edges
- ✓ Socio-cultural effects

- Differences in compositions of K-Lex and P-Lex
- Heads and modifiers

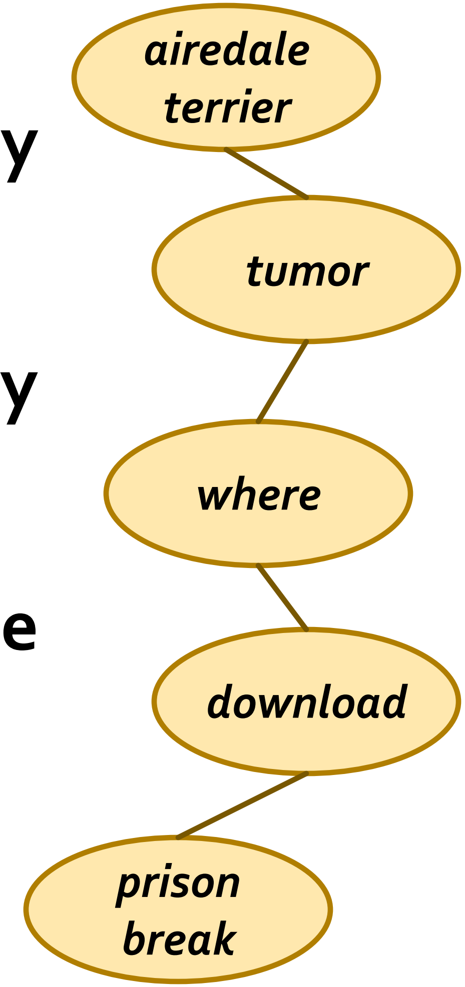
K-Lex (popular segments)	P-Lex (rarer segments)
how to	*matthew brodrick*
wiki	*accessories*
free	*police officer*
and	*who is*
in australia	*epson tx800*
videos	*star trek next gen*
real estate	*adams apple*
difference between	*harvard university*
windows xp	*leukemia*



Insights (2)

- ✓ K-Lex and P-Lex
- ✓ Higher mean shortest paths
- ✓ Less tight kernel
- ✓ More k-p edges
- ✓ Socio-cultural effects

- Higher mean shortest path in query networks
- Peripheral units can independently form queries
- More difficult to understand the context of a previously unseen unit
- High surprise factor



Insights (3)

- ✓ K-Lex and P-Lex
- ✓ Higher mean shortest paths
- ✓ Less tight kernel
- ✓ More k-p edges
- ✓ Socio-cultural effects

- **Kernel is less tightly coupled**
- **98% edges run between kernel and periphery, while intra-kernel edges dominate in English**
- **Socio-cultural factors govern kernel-periphery distinction (*lyrics*, *movies*, *adelaide* in K-Lex; *code*, *accessories*, *delhi* in P-Lex)**

Thank You!