Stream Characterization from Content

Allen Gorin

Human Language Technology Research

U.S. DoD, Fort Meade MD

a.gorin@ieee.org

Collaborators

Carey Priebe (JHU)

John Grothendieck (BBN)

Nash Borges

John Conroy

Glen Coppersmith

Rich Cox

Mike Decerbo

Dave Marchette

Alan McCree

Youngser Park

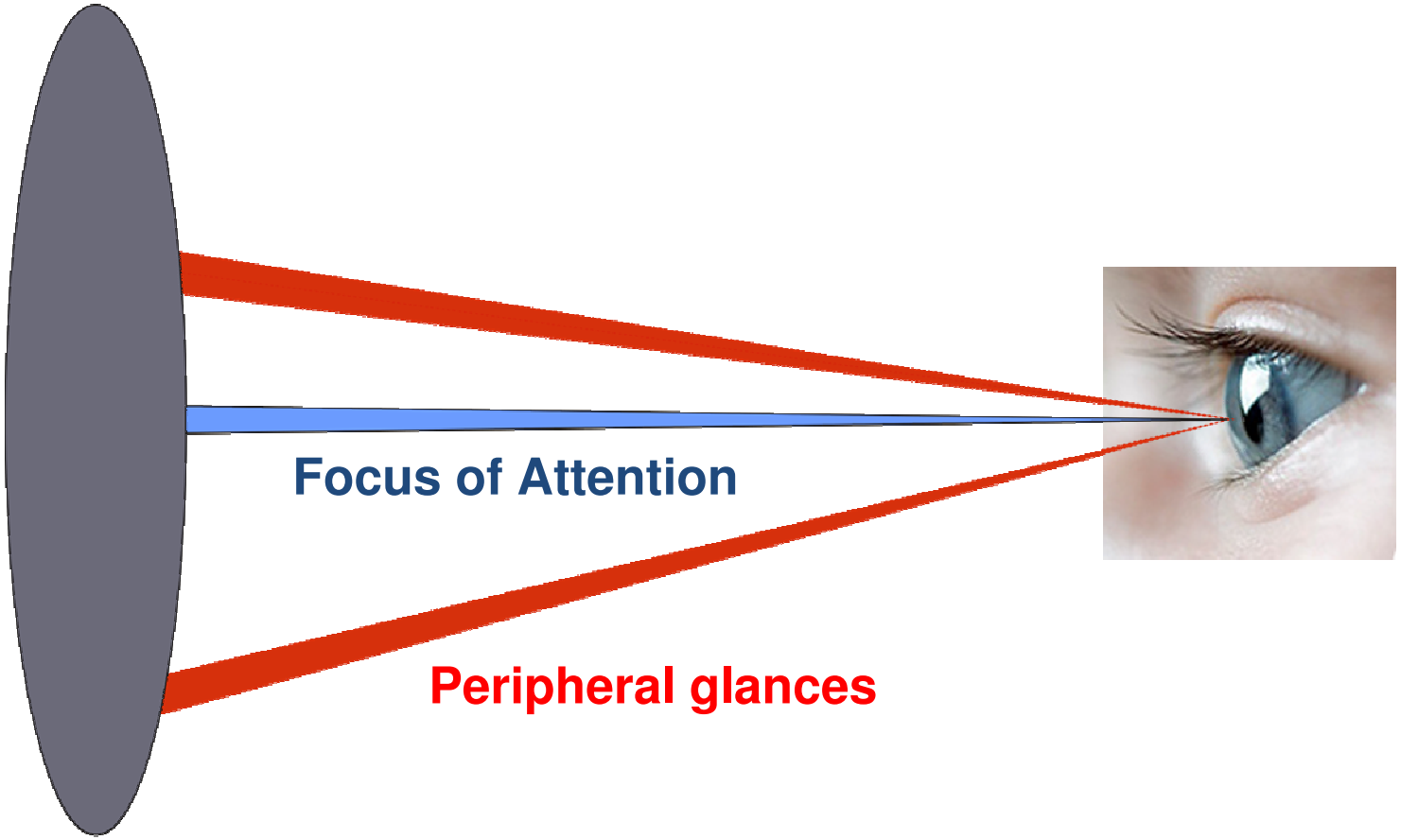
Alison Stevens

Jerry Wright

Outline

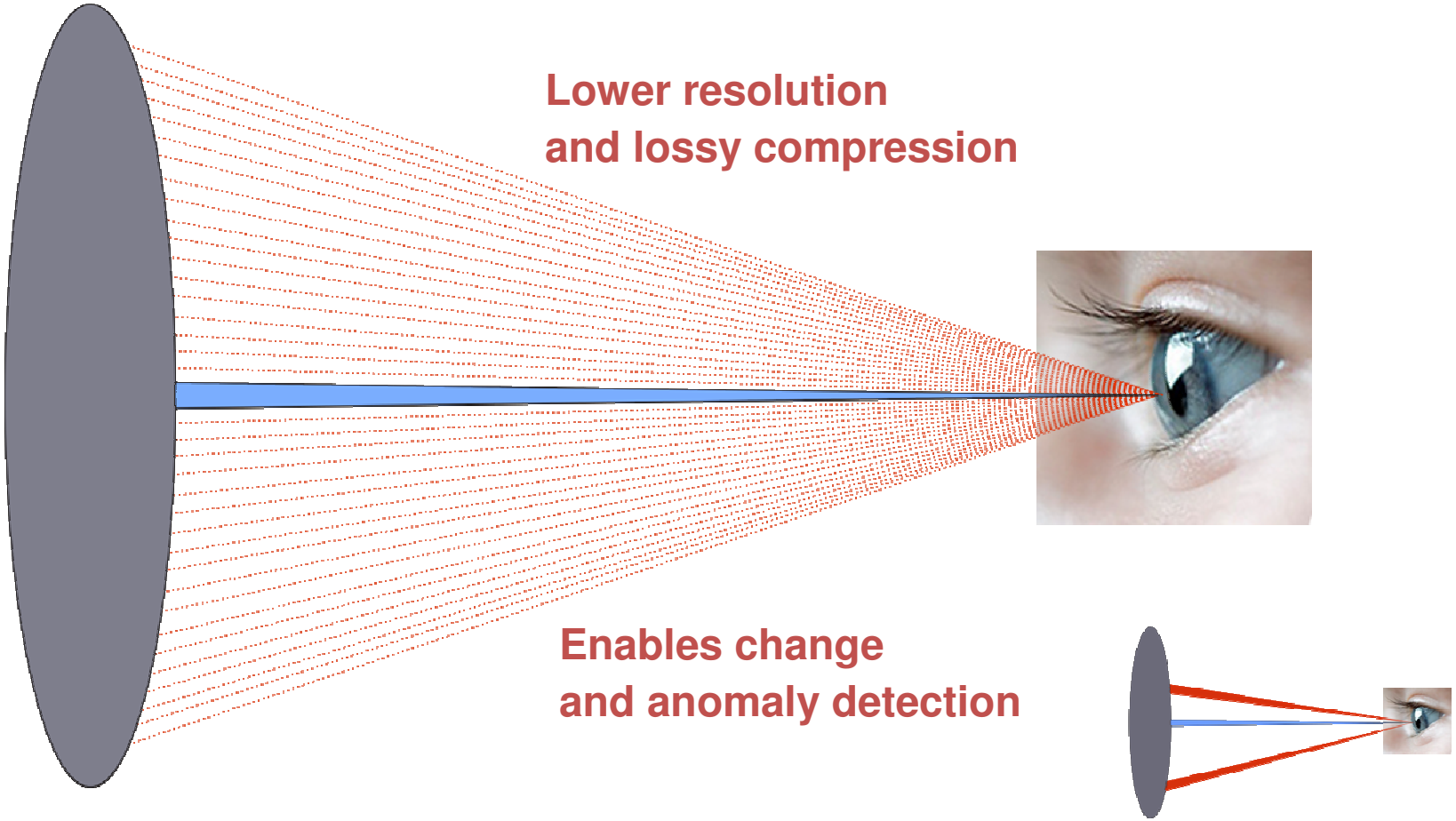
- **Motivation**
- HLT Research Issues
- Joint model of content in context
- Experiments on speech using Switchboard
- Experiments on text using Enron

Environmental Awareness



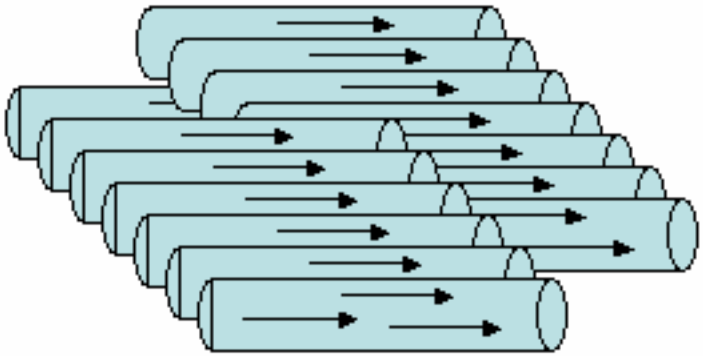
Environmental Awareness:

Focus of Attention *plus* Peripheral 'Vision'



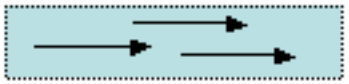
Coping with Information Overload

Data Streams and substreams



Bandwidth Reduction

Pick out the good stuff



Filter and Select

Boil it down

Stream Characterization

- Mature: External Metadata

- **Emerging: Metacontent**

  - *language*
  - *speaker*
  - *topic*

Analytic Questions

- Is the information environment stable?
  - describe environment
  - lossy compression
- Did something change?
  - Where? What?

Outline

- Motivation
- **HLT Research Issues**
- Joint model of content in context
- Experiments on speech using Switchboard
- Experiments on text using Enron

HLT Research Issues

- **Focus on stream statistics**
  - Rather than on individual documents
  - E.g. *Language Characterization* (McCree)
  - Classifier output is *biased* and noisy (Grothendieck)
  - Piece-wise stationary segments (Wright)
- **Content has associated meta-data**
  - Better living through *content in context*
  - Theory, simulations and experiments
  - with Priebe, Grothendieck, et al

Experimental Corpora

- **Enron corpus of emails**
  - 500K emails over 189 weeks from DoJ/CMU
  - 184 communicants
  - 32 topics as defined by LDC
- **Switchboard corpus of spoken dialogs**
  - 2500 topical dialogs
  - between pairs of 500 speakers
  - speaker demographics

Outline

- Motivation
- HLT Research Issues
- **Joint model of content in context**
- Experiments on speech using Switchboard
- Experiments on text using Enron

Joint model of content in context

- Consider a set of *communication events*

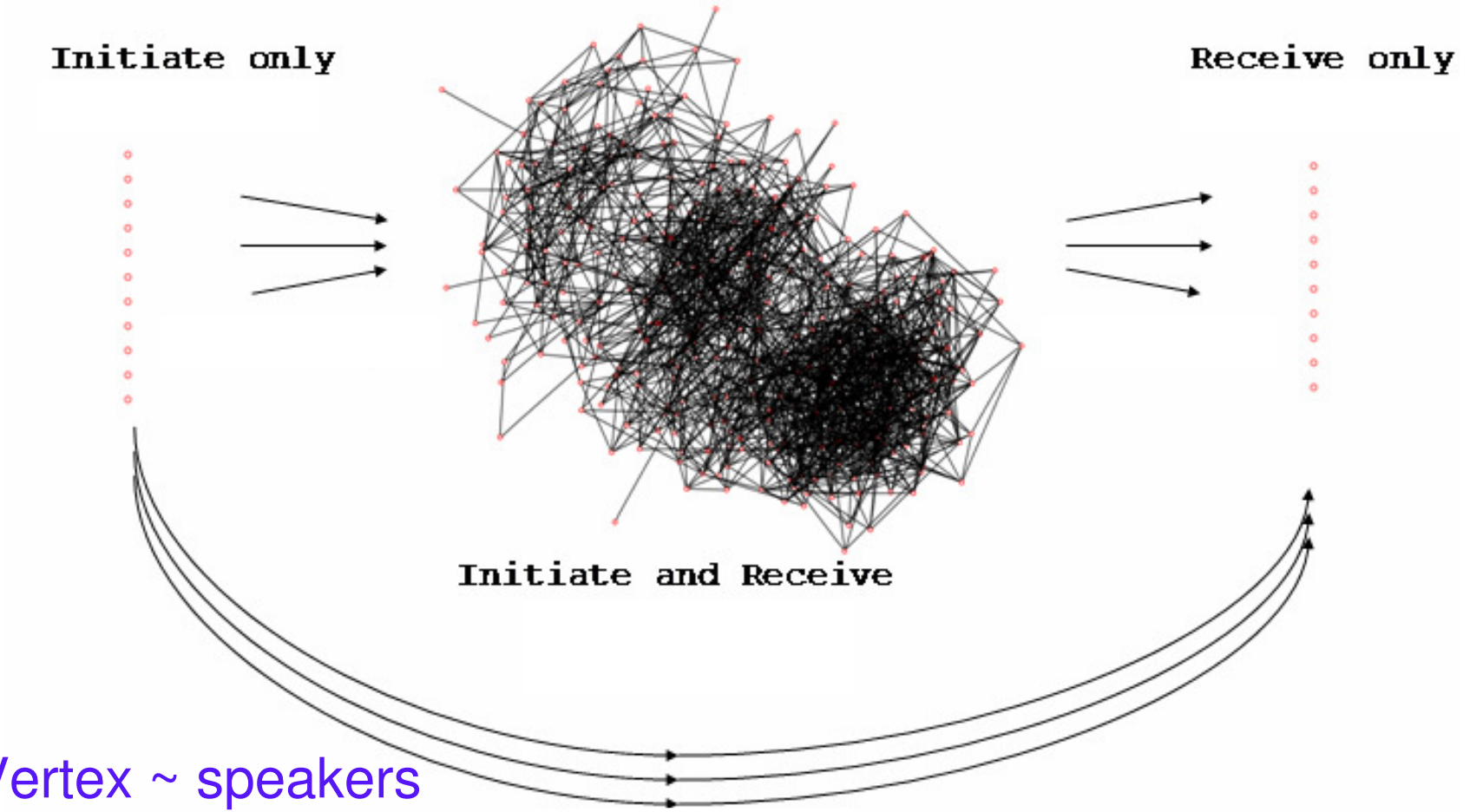
  $M = \{z_i = (u_i, v_i, t_i, x_i)\} \in \mathcal{M}$ with

- An event in M is $z_i \in V \times V \times R_+ \times \Xi$
  - representing *(to, from, time, content)*
- A time window defines a graph with content-attributed edges
- Attribution functions h_V and h_E to further color vertices and edges

Examples from Enron Corpus (high-dimensional and heterogeneous features)

Date	Time	Sender	Receiver	Sender's Rank	Topic
2001-01-02	04:15:00	steven.k	jeff.d	Vice President	(1) California Analysis
2001-02-09	13:49:09	louise.k	andy.z	President	(9) Daily Business
2001-02-16	21:06:00	drew.f	jeff.d	Vice President	(5) California Enron
2001-02-26	22:30:00	james.s	john.l	Vice President	(14) Energy Newsfeed
2001-03-01	07:54:00	diana.s	kate.s	Trader	(5) California Enron
2001-04-06	05:15:00	mike.g	john.l	Manager	(7) Newsfeed California
2001-04-16	06:12:00	richard.s	steven.k	Vice President	(9) Daily Business
2001-05-11	16:02:00	andy.z	john.l	Vice President	(11) Enron Online
2001-06-27	17:44:24	s..s	geoff.s	Vice President	(9) Daily Business
2001-09-05	14:36:53	geoff.s	louise.k	Director	(12) Enrononline Daily
2001-09-15	20:51:20	m..p	louise.k	Vice President	(12) Enrononline Daily
2001-10-04	14:19:16	john.l	louise.k	CEO	(11) Enron Online
2001-10-05	18:49:05	j..k	richard.s	Vice President	(9) Daily Business
2001-10-08	17:50:19	shelley.c	darrell.s	Vice President	(1) California Analysis

SwitchBoard Communications Graph



Vertex ~ speakers
Edges ~ dialogs

Joint Model of Content and Context via Attributed Graphs

- **Edge attributes**
  - Content-derived meta-data (a.k.a. *meta-content*)
  - E.g. topic id, ASR, turn-taking behavior
- **Vertex attributes**
  - *External meta-data* about speaker
  - E.g. demographics such as age, gender, education, …
  - *Graph-derived* meta-data
  - E.g. vertex degree ~ willingness to communicate

Outline

- Motivation
- HLT Research Issues
- Joint model of content in context
- **Experiments on speech using Switchboard**
- Experiments on text using Enron

Joint Model of Content and Context

- Random Attributed Graph
  - Provides a joint model of content and context
- In Switchboard
  - Content is an attribute of an edge (dialog)
  - Consider turn-taking behavior in the dialog
  - Context is an attribute of the vertices (speakers)
  - Consider age, education, gender of speakers
- Joint model enables inference of
  - *Unobserved* demographic distribution
  - From *observed* turn-taking behavior

Models of Turn-Taking Behavior

- Turn-taking behavior has predictive power
  - for speaker ID (Jones)
  - for speaker traits in meeting room data (Lakowski)
  - for social roles and networks (Pentland)
- *Joint model* of vertex, edge attributes and graph
  - social correlates of turn-taking behavior
  - Grothendieck and Borges
  - experiment to exploit joint distribution
  - *observed meta-content* (turn-taking)
  - *estimate unseen* demographic distributions

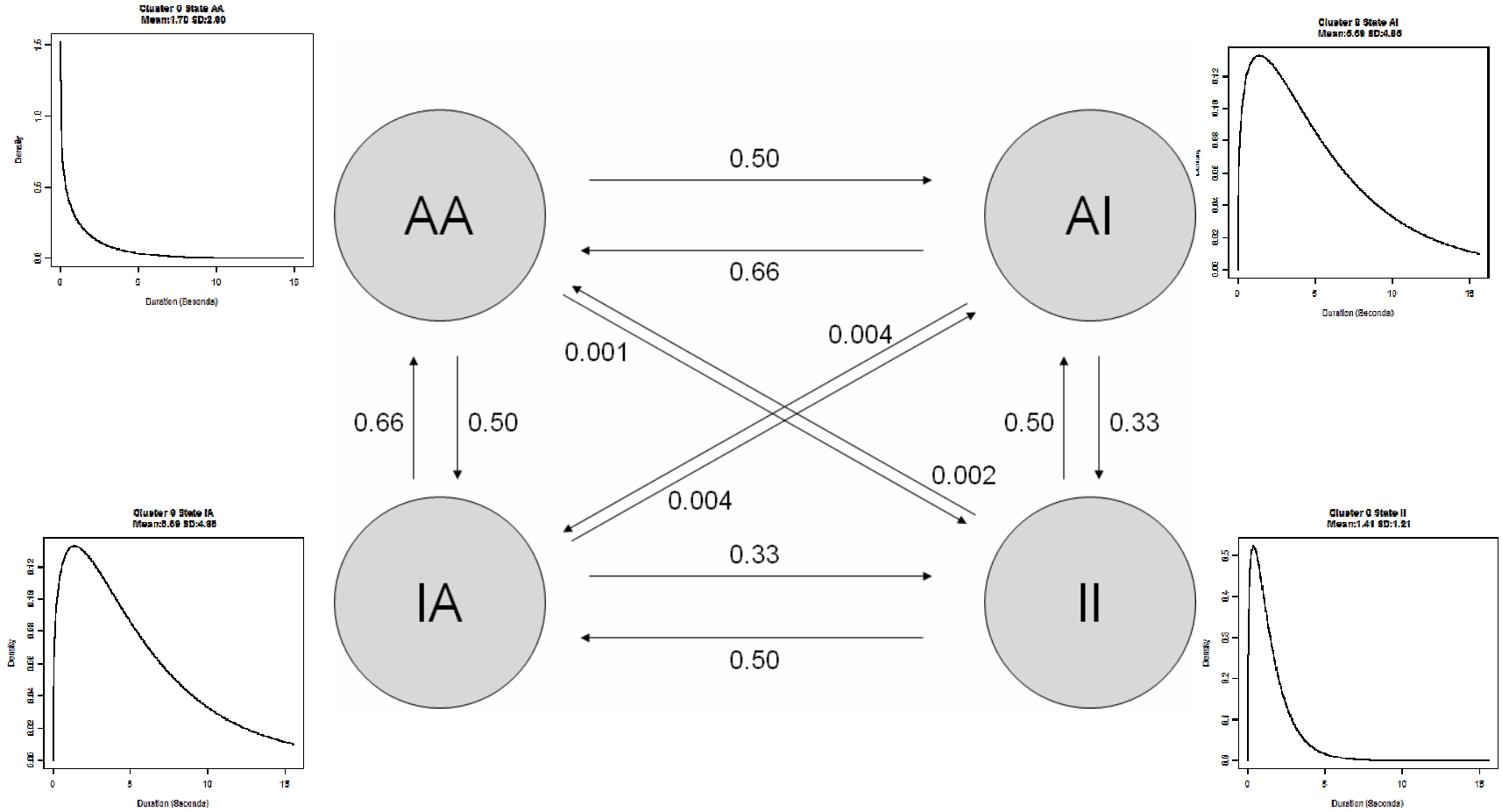
Turn-taking Behavior Model derived from SAD

Side 1: $S_1(t)$	I		A		I		A
Side 2: $S_2(t)$	A		I		A		I
Dialog State: S(t)	**IA**	**II**	**AI**	**AA**	**IA**	**II**	**AI**

A = active

I = inactive

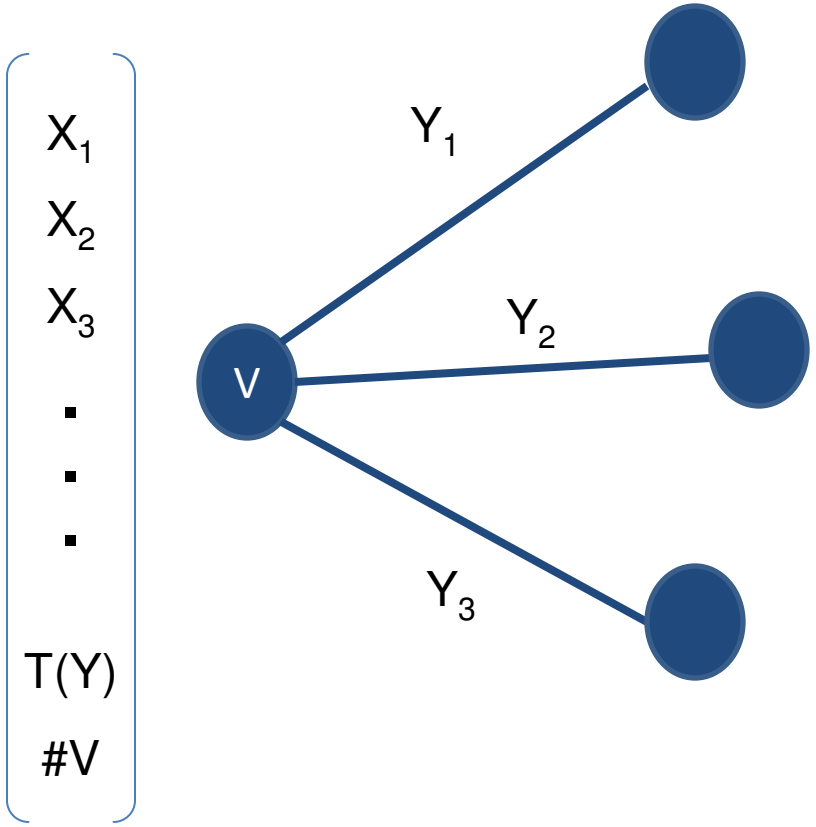
Semi-Markov Model of Turn-Taking Behavior



Latent Classes of Turn-Taking Behavior

- Train turn-taking model from Switchboard corpus
- First-order partition via *divisive clustering*
  - E.g., Style 0 has more and longer **II** (both silent)
  - E.g., Style 1 has more and longer **AA** (both active)
- Classify each dialog as style 0 or 1
  - Edge attribute (meta-content)
- Classify each speaker as having style 0 or 1
  - Vertex attribute induced from edge attributes

Enriching vertex attributes with edge meta-content and graph meta-data



- ***X*** = external meta-data on speaker v
- ***Y*** = conversation turn-taking style
- ***T(Y)*** = turn-taking style of speaker v
- **#V** = number of conversations including speaker v

Experimental Evaluation

- E.g., overall ratio of male:female is 1:1
  - speakers with *TT style 0 have ratio 2:1*
- Have joint distribution of content and context
  - exploit *observed content* (turn-taking behavior)
  - to *estimate unobserved context* (demographic mix)
- *Experiment*: create speaker sets with mixture proportion *v* of style 0, for *v* in [0,1]
- Result: across all mixtures *v* of styles,
  - predict proportions of age, education, gender, …
  - yields RMS error ~ 0.1

Classic Problems in DSP

- Estimate characteristic parameters
  - Oppenheim (1975)
- To detect a signal in background noise
  - Van Trees (1968)
- Motivates initial focus on change/anomaly detection

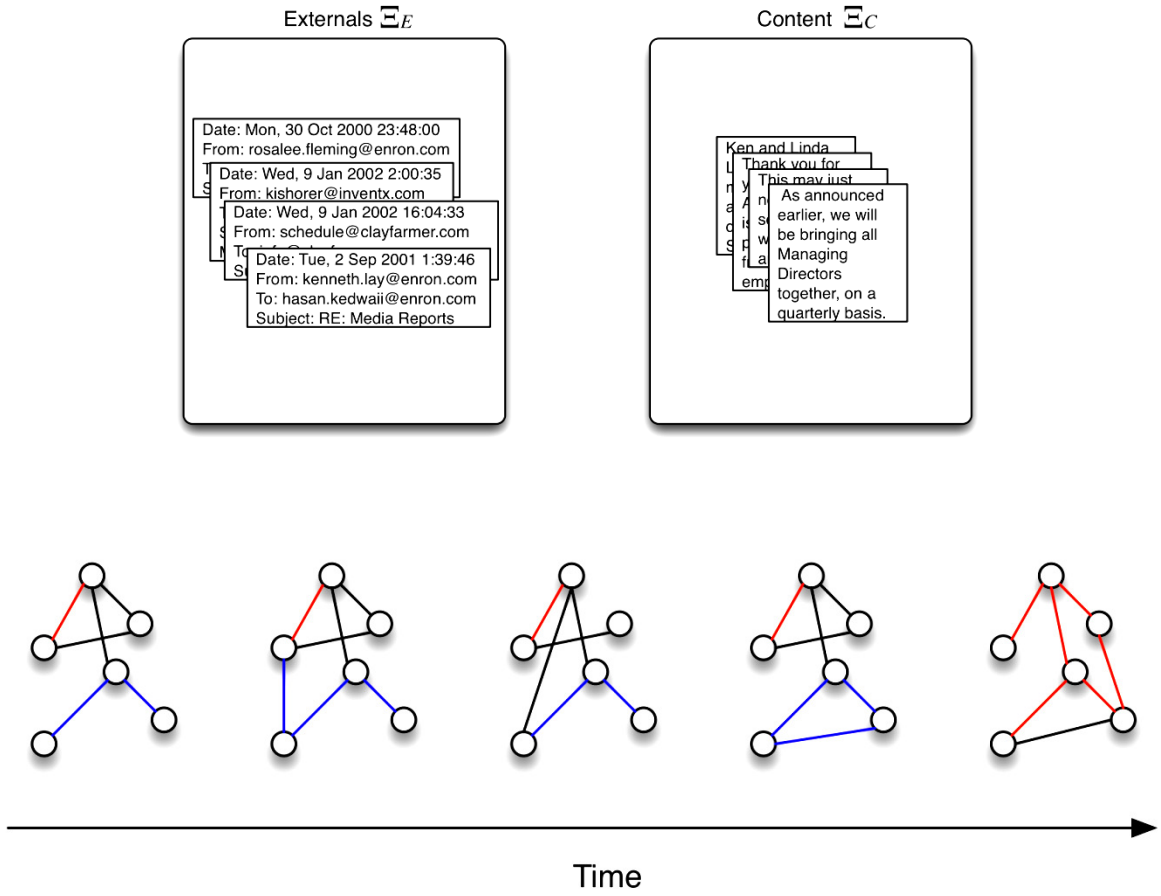
Better Living through Content in Context

- *Information Exploitation* = statistical inference
- *Better* = more powerful statistical test
  - *for* change/anomaly detection
- Some results to date
  - Theorem that joint can be more powerful
  - Simulation experiments
  - Proof-of-concept experiment on Enron Corpus

Outline

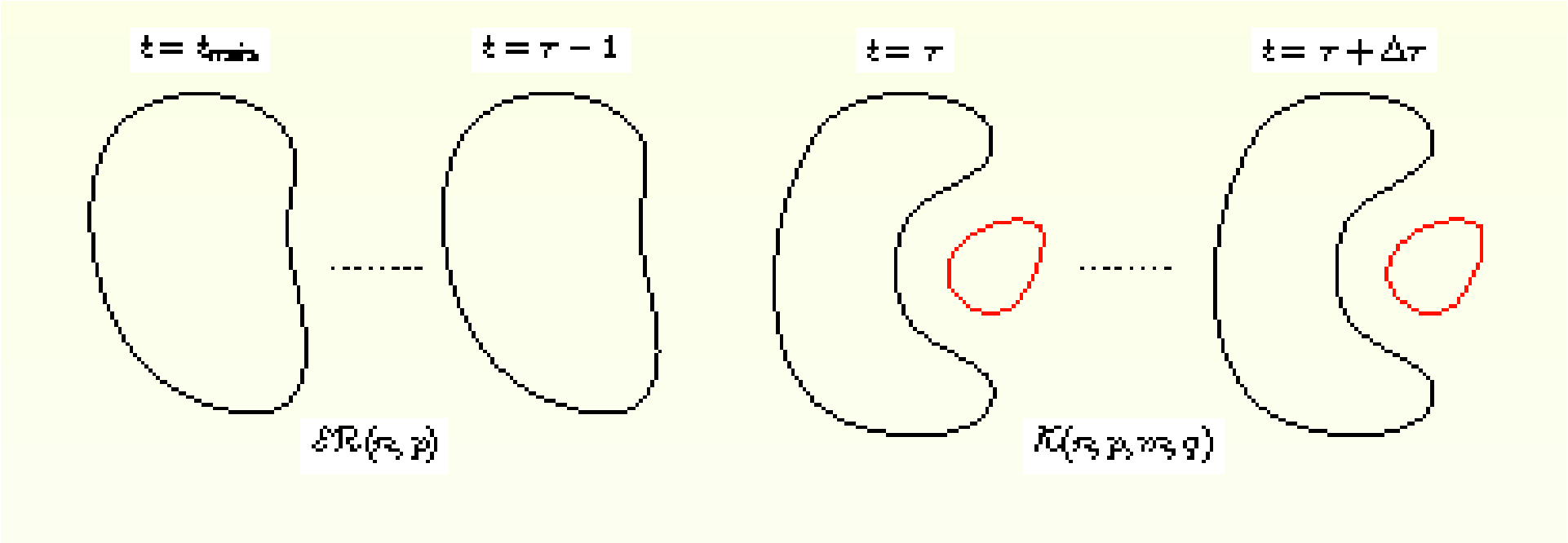
- Motivation
- HLT Research Issues
- Joint model of content in context
- Experiments on speech using Switchboard
- **Experiments on text using Enron**

Time Series of Attributed Graphs



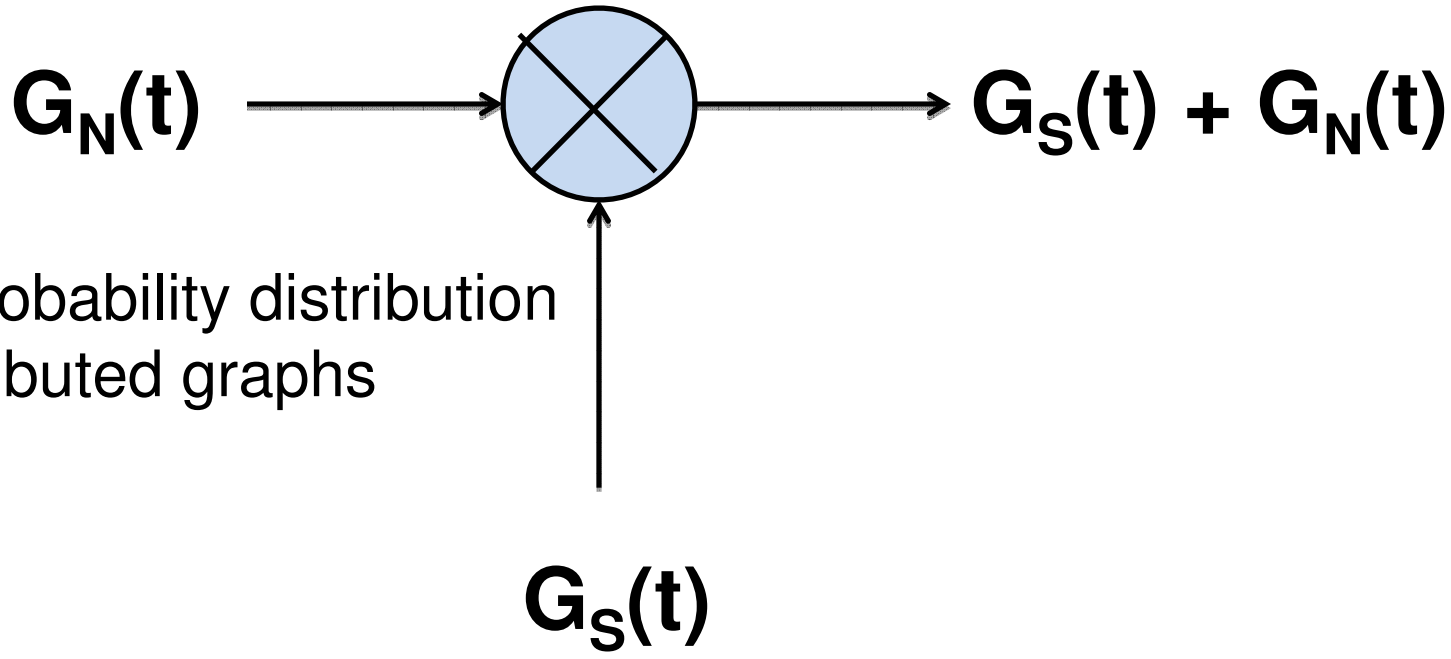
Generated from observations of some random attributed graph?

Change detection in a time series of Graphs



Homogeneous **Anomalous Chatter Group**

Detecting 'Signal' in 'Noise' - models and theory



Random Attributed Graphs

- Let's work through an example with a very simple model of content and context
- Existence of an edge between two vertices is IID Bernoulli with probability **p**
- Content topic (on each edge) is IID Bernoulli with probability **θ**
- Change detection via testing candidate anomaly (alternative) versus history (null)

Null Hypothesis (noise): an attributed Erdos-Renyi Graph

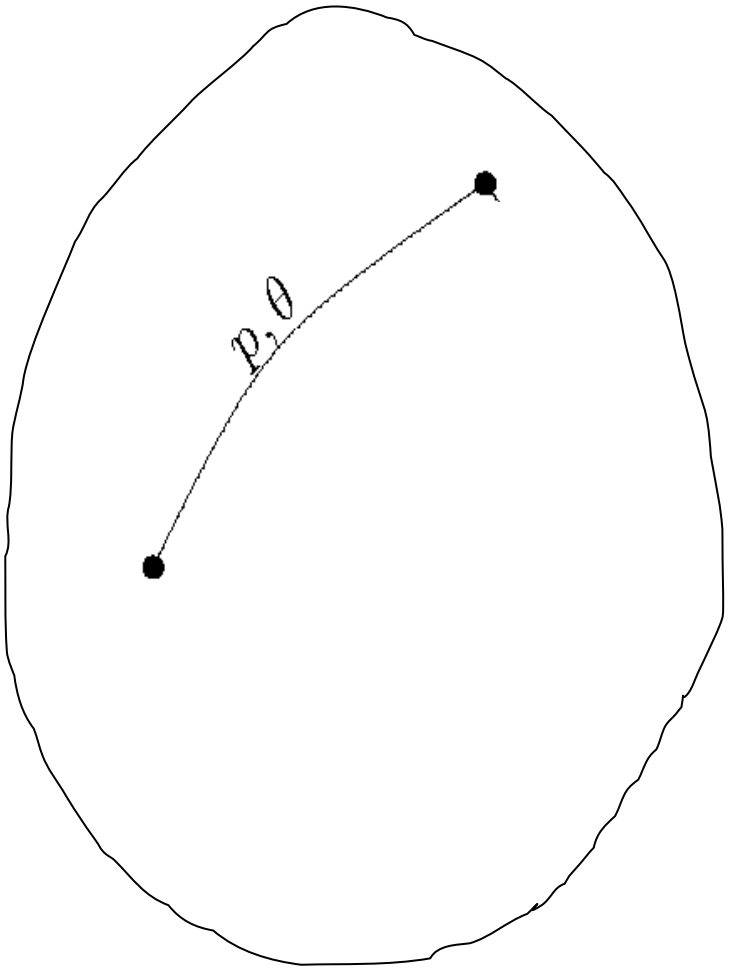
Random Graph ERC(N, p, Θ)

N = # vertices in the graph

p = probability of an edge

Each edge labeled

- with topic 0 or 1
- with Θ = probability of topic 1



Alternative Hypothesis (noise + signal): an ERC subgraph with different parameters

Random Graph

***K*(N,p, Θ, M, q, Θ')**

N = # vertices in whole graph

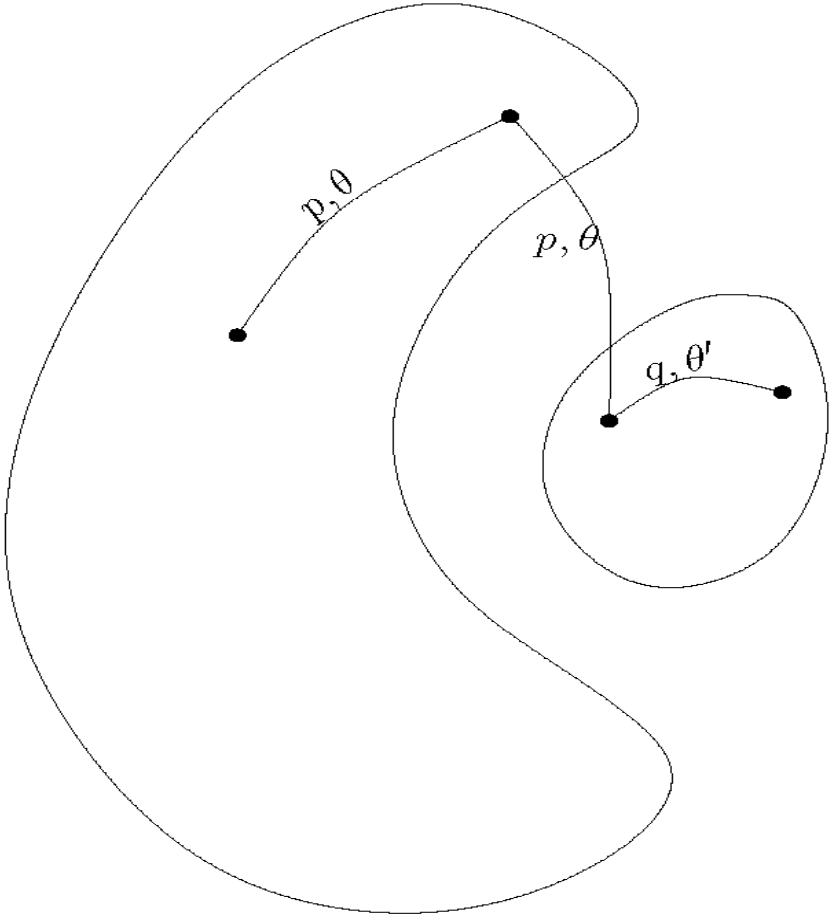
p = prob(edge) in kidney

Θ = topic parameter in kidney

M = # vertices in egg

q = prob(edge) in egg

Θ' = topic parameter in egg



Theorem

A statistical test based on fusion of externals and content *can* be more powerful than a test based on externals alone or content alone.

(Grothendieck and Priebe)

Proof by Construction

- T_G = # of graph edges
- T_C = # of graph edges attributed with topic 1
- $T = 0.5\ T_G + 0.5\ T_C$
- Test for change from homogeneous null graph:
  - Power of test based upon T_G is β_G
  - Power of test based upon T_C is β_C
  - Power of test based upon T is β
- For tests with false alarm rate $\alpha = 0.05$,
  - gray-scale plot of power difference $\Delta = \beta - \max(\beta_G, \beta_C)$

Power Difference: $\Delta = \beta - \max(\beta_C , \beta_G)$

Δ(Θ', q) depends on the parameters of the anomalous chatter group

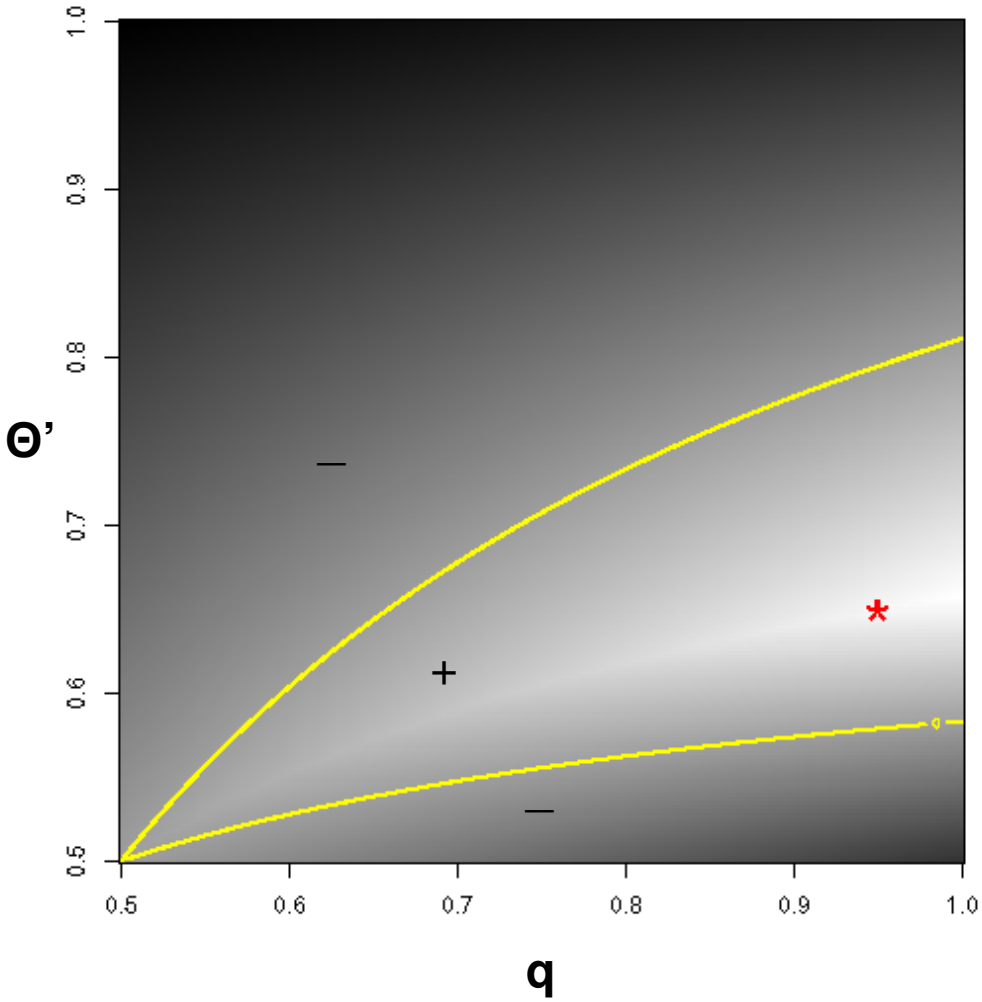
p = 0.5

Θ =0.5

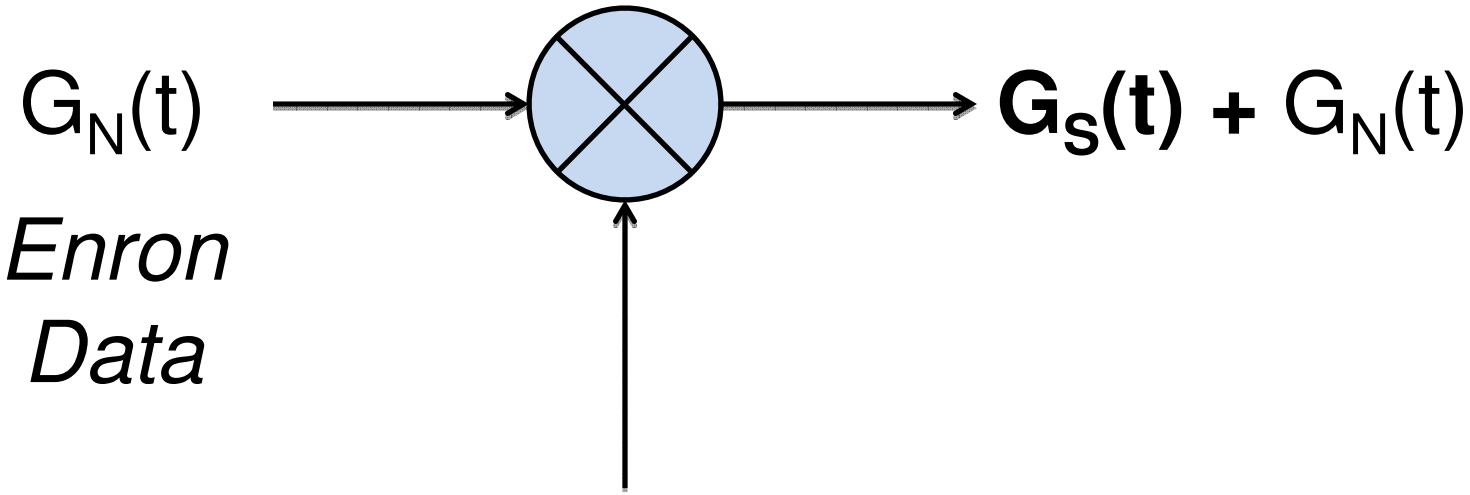
q = subgraph connectivity

Θ' = subgraph topic

Grayscale = Δ (Θ' , q)

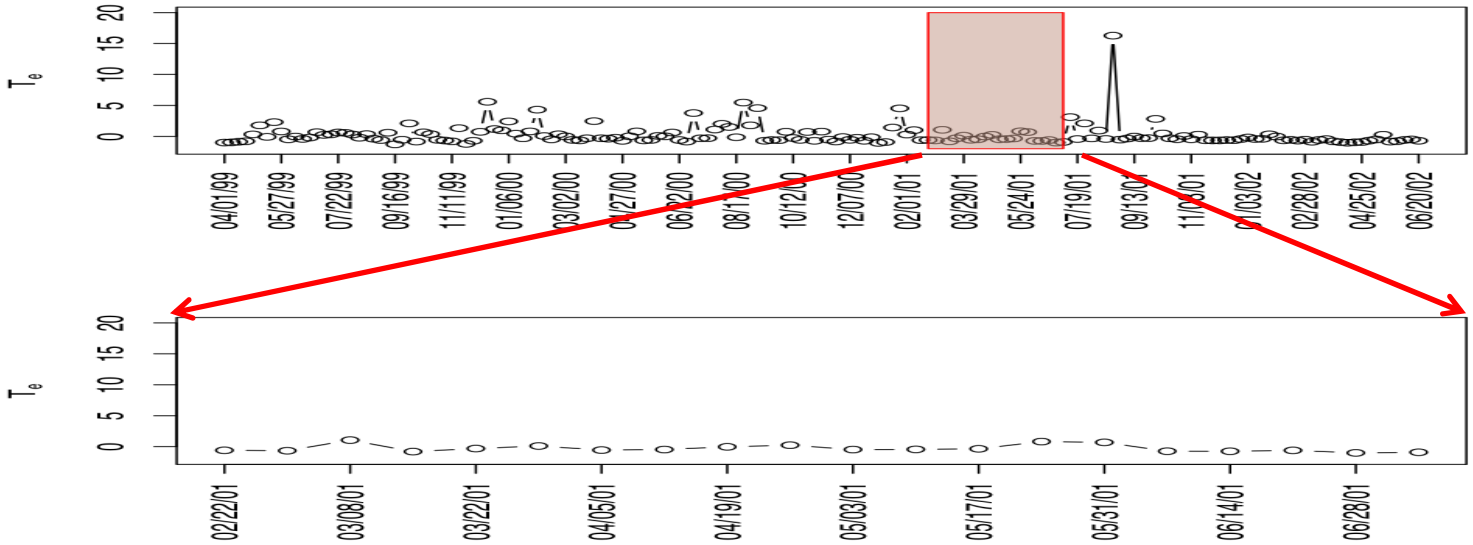


Detecting 'Signal' in Empirical 'Noise'



Enron Experiment

- Select a stationary region of test statistics for Enron
- Estimate empirical null $G_N(t)$ from that region
- Add 'signal' via model $\mathbf{G_S(t)}$ which injects egg
- Similar experimental results on power difference!



Conclusions

- Better living through *content in context*
  - modeled via *random attributed graphs*
- Better = *more powerful statistical inference*
- Joint model of content and context can be more powerful for many inference tasks
- *Theorem* for change/anomaly detection
- *Proof of Concept Experiments*
  - Inference of demographics from turn-taking behavior
  - Change/Anomaly detection
  - On Switchboard and Enron corpora

Acknowledgements

- Charles Wayne for
  - insights into communication graphs
- Deb Roy for
  - insights into content in context
- Sandy Pentland for
  - insights into social networks and communications

Some References

- **Random Attributed Graphs for Statistical Inference from Content and Context,** Gorin, Priebe and Grothendieck, Proc. ICASSP 2010.
- ***Statistical Inference on Random Graphs: Fusion of Graph Features and Content,*** Grothendieck, Priebe, and Gorin, Computational Statistics and Data Analysis (2010)
- ***Statistical Inference on random attributed Graphs: Fusion of Graph Features and Content: An Experiment on Timeseries of Enron Graphs,*** Priebe et al, Computational Statistics and Data Analysis (2010).
- ***Social Correlates of Turn-taking Behavior,*** Grothendieck, Gorin, and Borges, [ICASSP 2009] , [full paper submitted]
- ***Towards Link Characterization from Content: Recovering Distributions from Classifier Output***, Grothendieck and Gorin , IEEE Transactions on Speech and Audio, May 2008
- ***CoCITe – Coordinating Changes in Text,*** Wright and Grothendieck, *to appear, IEEE Trans. on Knowledge and Data Engineering*