Quantitative Horizon Scanning in Co-Authorship Networks

David J. Marchette

Joint Statistics Meetings
August 5, 2010

Collaborators

- Carey Priebe (JHU)
- Jeffrey Solka (NSWC)
- Avory Bryant (NSWC)

Outline

Outline

Quantitative Horizon Scanning

- Using scientific literature (publications, patents, etc) QHS attempts to:
  1. Determine areas in which breakthroughs are likely;
  2. Identify individuals or small groups working (or privy to work) in these areas;
  3. Quantify the likelihood of breakthrough in different areas;
  4. Quantify the future impact of scientific work;
  5. Suggest areas of research to mitigate technological surprise.
- Keys to performing QHS:
  - Data: bibliometrics.
  - Algorithms: text and network analysis.
- Statistical inference is the goal.

Guiding Principles

1. Technological surprise often involves the (unanticipated) fusion of ideas from disparate subject areas.
2. Identification of individuals or small groups working (or privy to work) in disparate subject areas is a quantitative horizon scanning inference task which can help mitigate technological surprise.

Topics

- We need to define “disparate subject areas” or topics.
- Usually, this is done by clustering the documents.
- Alternatively, one could use human-generated clusters.
- Question: can we use keywords as surrogates for clusters?

Surprise

- Given:
  - a definition of topic;
  - a measure of similarity between topics;
- We define surprise in terms of this similarity:
  - People working in very dissimilar topics, which are rarely seen together, are likely to produce surprising results.
- The last piece is a way of measuring “rarely” and “seen together”.
- We do this via privy analysis.

Outline

Privy Analysis

- We utilize the idea of "privy-to-information" to define a statistic to measure the likelihood that a scientist knows about a body of literature, a scientific discipline, etc.
- Illustrative example: An author is
  - 0-privy to a topic if she has written on the topic.
  - 1-privy to a topic if she has co-authored a paper with another author who has written on the topic.
  - ...
  - k-privy to a topic if she has co-authored a paper with another author who is $(k-1)$-privy.
- This provides a way of describing how "far away" an author with expertise in topic $\mathcal{T}_1$ is from knowledge about topic $\mathcal{T}_2$.

Privy Analysis: Data

- Given $\mathcal{D}$, a collection of text documents from the scientific literature.
- Define a collection of functions on $\mathcal{D}$:
  - f_a which extracts document authors,
  - f_c which extracts document country affiliations,
  - f_i which extracts document institution affiliations,
  - f_s which extracts document subject identifications,
  - f_k which extracts document keywords,
  - f_A which extracts document abstract,
  - . . .
  - f_t which extracts document time stamp (year).

Privy Analysis: Data

- Thus each $d \in \mathcal{D}$ has associated with it
  - an author collection $f_a(d)$,
  - a country affiliation collection $f_c(d)$,
  - a institution affiliation collection $f_i(d)$,
  - $\cdots$, and
  - a time stamp $f_t(d)$.
- These are the raw data used for privy analysis.

Privy Analysis: Author Attributes

- Let $\mathcal{A} = \cup_d f_a(d)$ be the overall collection of authors active in $\mathcal{D}$.
- Let $\mathcal{C} = \cup_d f_c(d)$ be the overall collection of countries active in $\mathcal{D}$.
- Let $\mathcal{I} = \cup_d f_i(d)$ be the overall collection of institutions active in $\mathcal{D}$.
- From this, we can associate *author attributes* with each $a \in \mathcal{A}$, such as country affiliation $g_c(a) \subset \mathcal{C}$, institution affiliation $g_i(a) \subset \mathcal{I}$, etc.

Privy Analysis: Social Network

- Let $G = (\mathcal{A}, E)$ be a social network on $\mathcal{A}$.
- For instance, we have available $G = (\mathcal{A}, E; \mathcal{D})$, the co-authorship graph induced by $\mathcal{D}$, where $uv \in E \iff \exists d \in \mathcal{D}\ s.t.\ \{u, v\} \subset f_a(d)$.
- We can augment G with edges based on institution affiliation, etc.
- We can also incorporate information external to $\mathcal{D}$, such as graduate student/postdoc connections, etc.
- NB: A hypergraph representation may be more natural, but it is unclear whether such a representation would be more useful.

Privy Analysis: Colored Social Network

- Given $\ell \subset \mathcal{D}$ (e.g. documents defining a topic), let G_ℓ be the social network $G = (\mathcal{A}, E)$ where edges $uv \in E$ are ℓ-colored based on co-authored documents in ℓ.
- $uv \in E$ is ℓ-colored $\iff \exists x \in \ell \; s.t. \; \{u, v\} \subset f_a(x)$.
- For $a \in \mathcal{A}$, let $d_\ell(a)$ denote graph distance in G_ℓ from vertex a to an ℓ-colored edge;

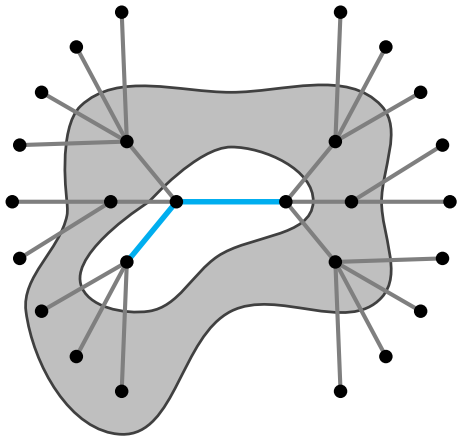
$$d_\ell(a) = \min_k \{ \exists \, \ell\text{-colored edge incident to a vertex } v \in N_k[a] \}.$$

Privy Analysis: Privy Sets

Given $b \in \mathbb{Z}_+$, author subset $\mathcal{Y} \subset \mathcal{A}$, and document subset $\ell \subset \mathcal{D}$, consider

$$A(b, \mathcal{Y}, \ell) = \{a \in \mathcal{Y} \; s.t. \; d_\ell(a) \leq b\}.$$

Definition: The set $A(b, \mathcal{Y}, \ell)$ represents the collection of authors in $\mathcal{Y}$ who are b-privy to the ℓ-subject.



Privy Analysis: Odds

- Given $b \in \mathbb{Z}_+$, author subset $\mathcal{Y} \subset \mathcal{A}$, and document subsets $\ell_1, \ell_2, \mathcal{D}' \subset \mathcal{D}$ with $\ell_1 \cup \ell_2 \subset \mathcal{D}'$, consider

$$p^{\mathcal{Y}} = |A(b, \mathcal{Y}, \ell_1) \cap A(b, \mathcal{Y}, \ell_2)| / |A(b, \mathcal{Y}, \mathcal{D}')|.$$

  (Note that $p^{\mathcal{Y}}$ depends on $b, \ell_1, \ell_2, \mathcal{D}'$ as well as $\mathcal{Y}$; we suppress this dependence for notational convenience.)

Definition: The quantity $p^{\mathcal{Y}}/(1 - p^{\mathcal{Y}})$ represents the odds (with respect to subcorpora $\mathcal{D}'$) that $\mathcal{Y}$-authors are b-privy to the (ℓ_1, ℓ_2)-interface.

Privy Analysis: Log-Odds Ratio

- Let $b \in \mathbb{Z}_+$, $\mathcal{Y}_1, \mathcal{Y}_2 \subset \mathcal{A}$, and ℓ_1,ℓ_2,$\mathcal{D}' \subset \mathcal{D}$ with $\ell_1 \cup \ell_2 \subset \mathcal{D}'$.

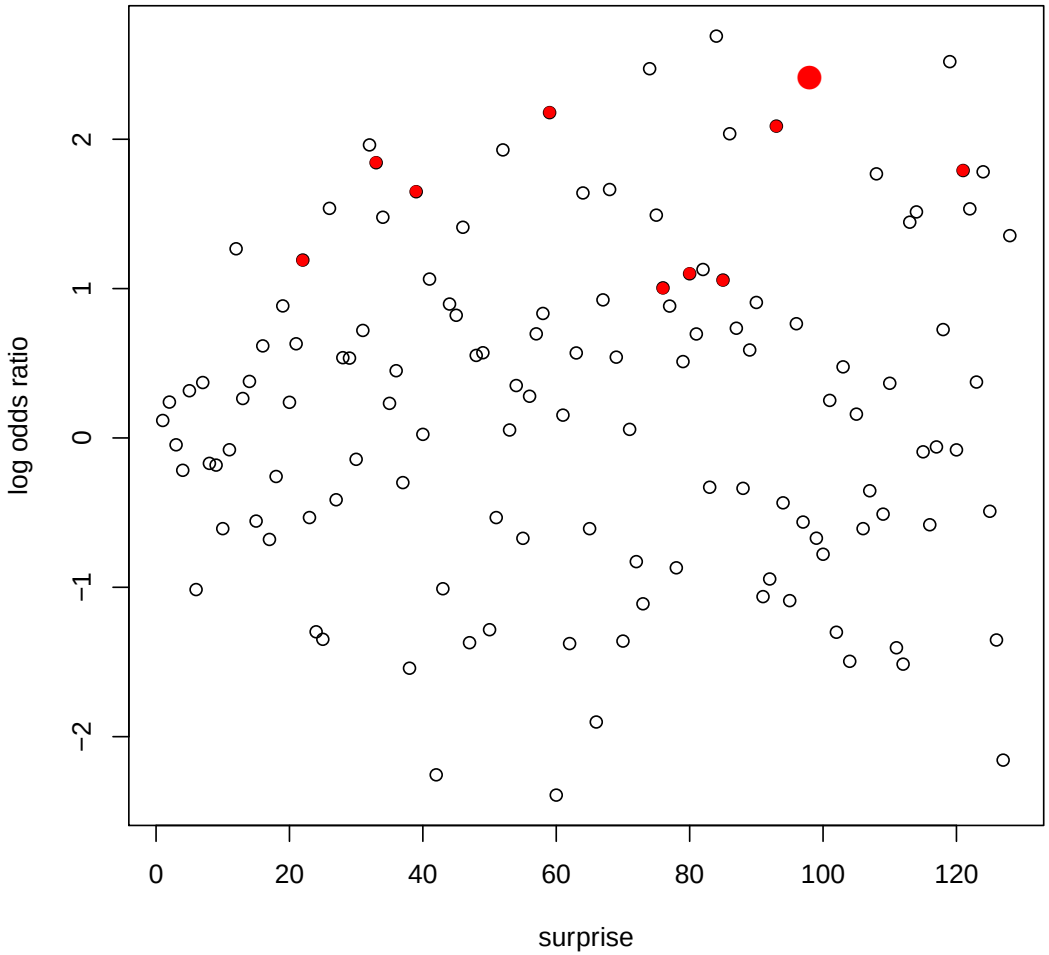
Definition: The statistic

$$\widehat{L}(b, \mathcal{Y}_1, \mathcal{Y}_2, \ell_1, \ell_2, \mathcal{D}') = \ln\left(\frac{p^{\mathcal{Y}_1}/(1-p^{\mathcal{Y}_1})}{p^{\mathcal{Y}_2}/(1-p^{\mathcal{Y}_2})}\right)$$

is the log-odds-ratio for differential b-collaborative potential at the (ℓ_1, ℓ_2)-interface for $\mathcal{Y}_1$ with respect to $\mathcal{Y}_2$.

- Large, statistically significant $\widehat{L}$ between high surprise topics, are candidates for detections.

Privy Analysis



Outline

Issues

- Author disambiguation.
- Topic identification and tracking.
- Missing data:
  - How complete is our collection of scientific documents?
  - Do we have all the meta-data, and have we extracted it correctly?
  - We cannot measure what the authors are reading – can conference attendance be used?
- Different publication rates in different communities.
- Patents vs journals.

Questions

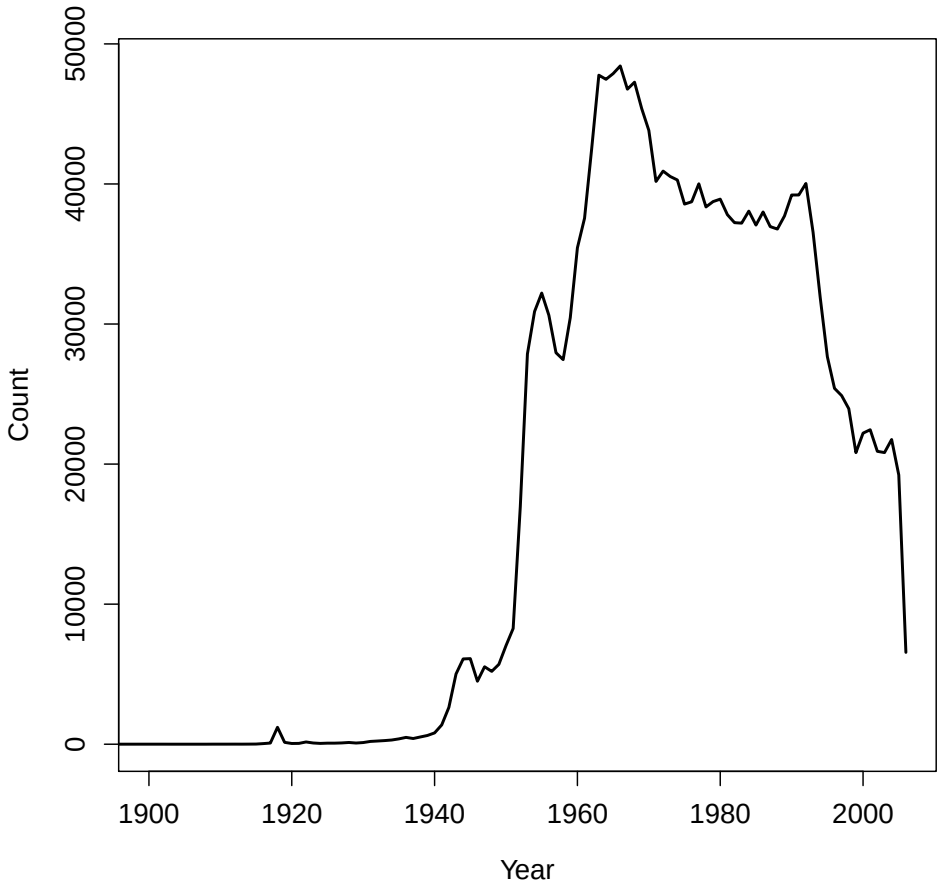
- What are the right time windows? Do scientists forget?
- Can privy analysis be used to infer technological surprise:
  - Do authors reduce their privy numbers in time?
  - Specifically: does 1-privy at time t tend to result in 0-privy at time $T > t$?
- Are keywords a good surrogate for topics?

Outline

DTIC Data

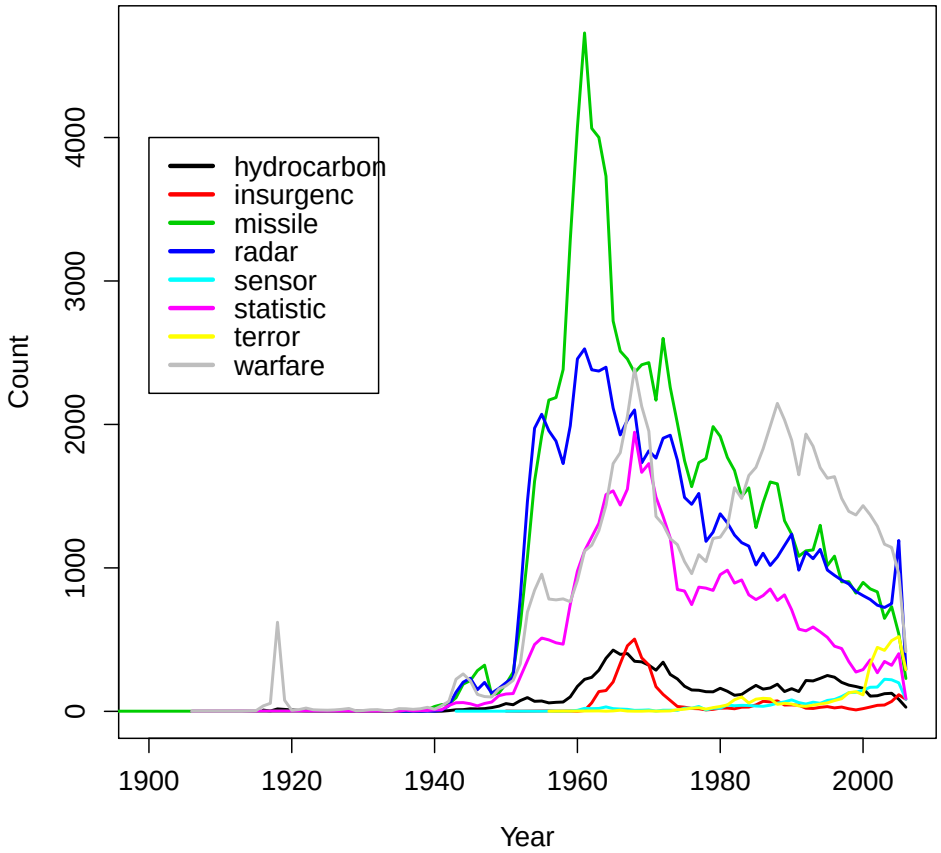
- Defense Technical Information Center.
- Technical reports from about 1900 through 2006.
- We consider articles published between 1980 and 1990 on several topics.
- Topics defined in this study by the keywords provided by the authors.
- We consider the privy value of an author to a topic across time.
  NB: for each time, we only consider papers during that time, ignoring the past.

Number of Publications



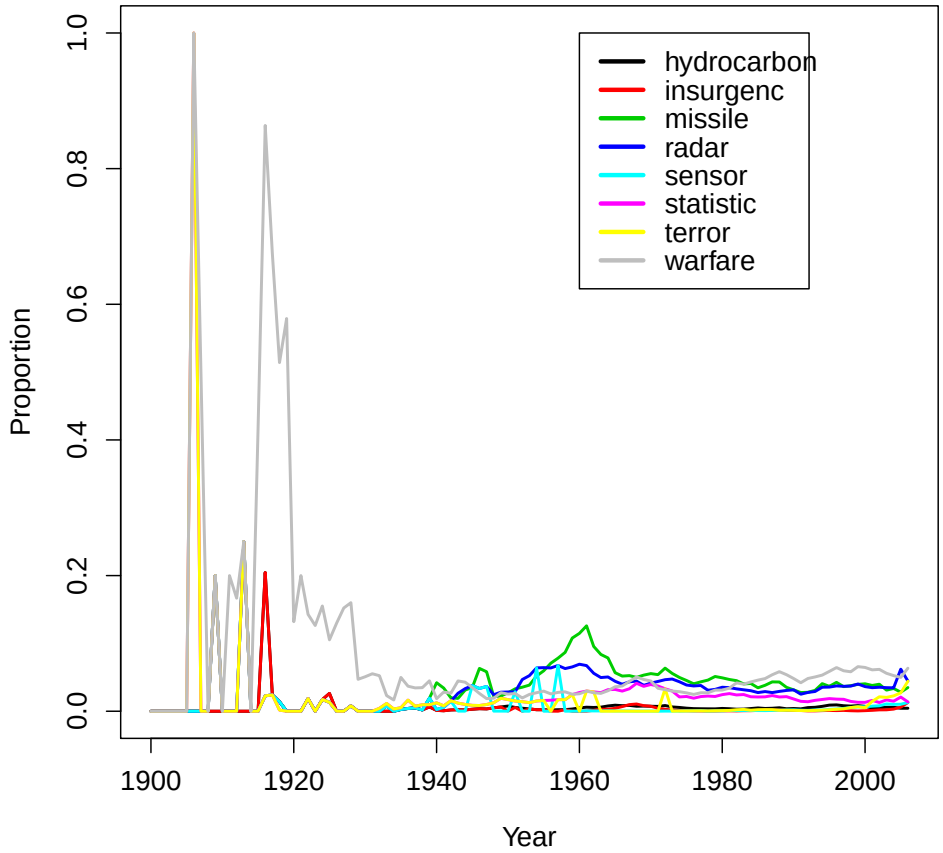
Publication rates falling: funding agency changes?

Topics: Number of Publications

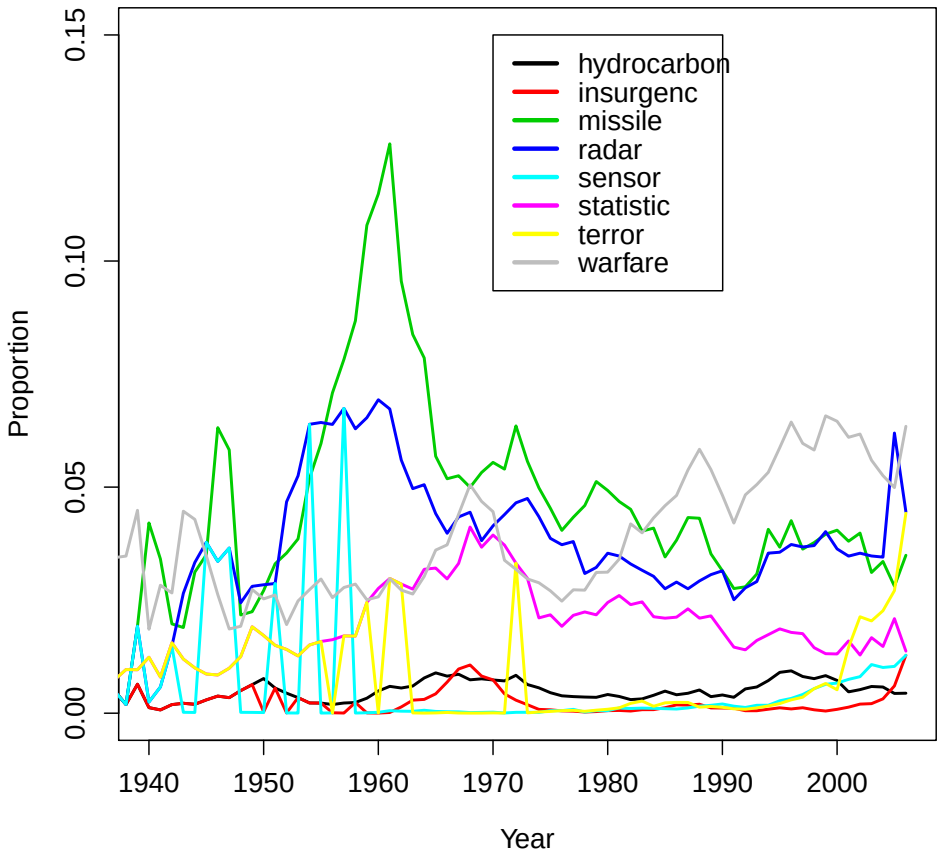


Keyword counts drop in time – tracking pub. rates.

Topics: Proportion of Publications



Topics: Proportion of Publications: 1940–2006

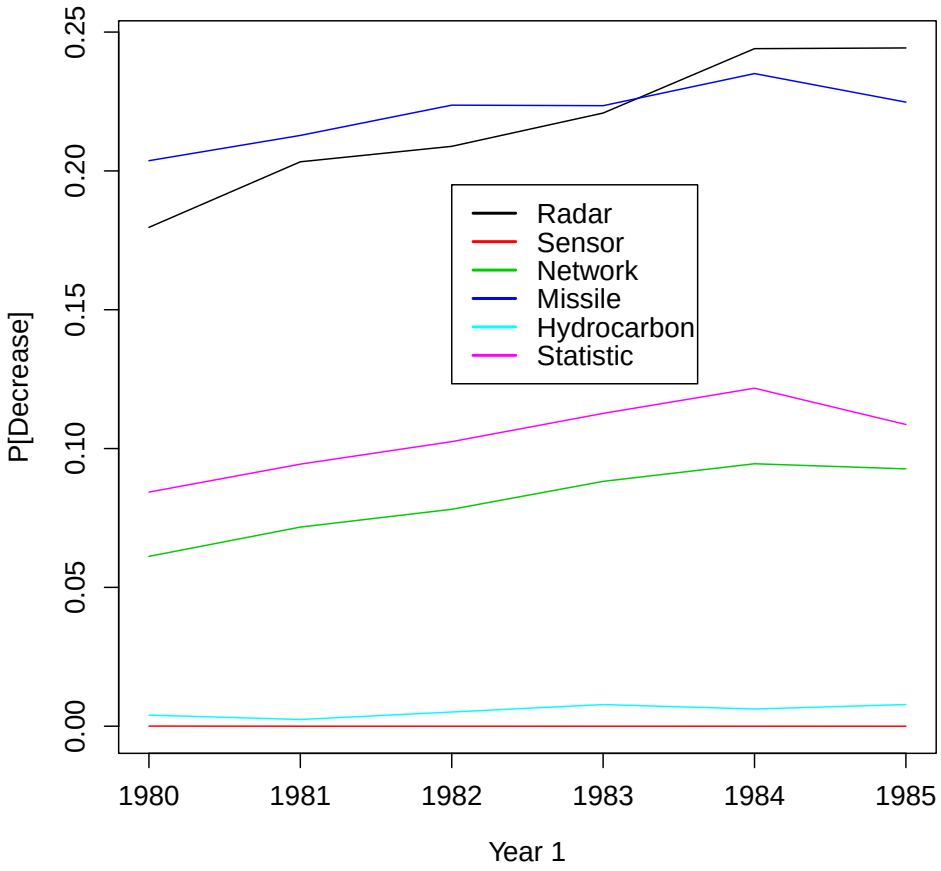


Some evidence generic keyword use drops off.

Does b Decrease?

- Experiment:
  1. For $t_1 \in \{1980, \ldots, 1984\}$ and a collection of keywords compute priviness for each author/keyword pairing in co-authorship graph for year t_1.
  2. For co-authorship graph computed over the next 5 years, $\{t_1 + 1, \ldots, t_1 + 5\}$, compute priviness for each author/keyword pairing.
  3. Return proportion in which priviness decreased.
- Evaluates rates at which authors decrease their "distance" to topics.

Decreasing *b*

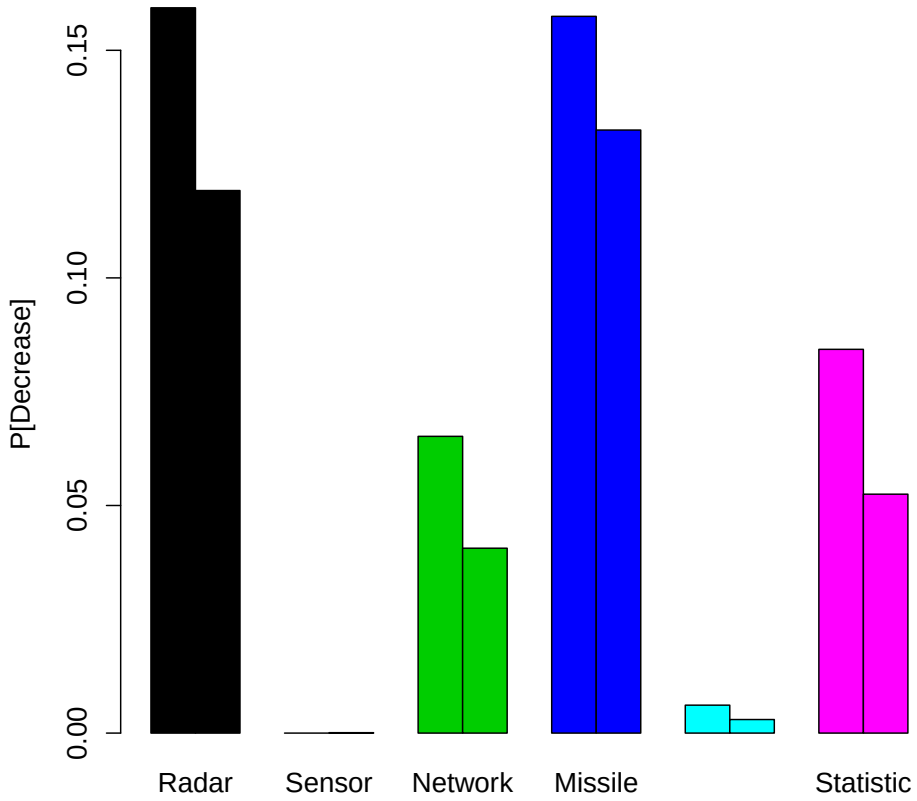


Comparing priviness at t_1 against the next 5 years.

Can *b* Increase?

- Same experiment, only taking $t_1 = \{1980, \ldots, 1984\}$ and $t_2 = \{1985, \ldots, 1989\}$.
- Since we are using disjoint windows, and not adding to previous graph, priviness is computed independently in each graph, for each keyword.
- In this scenario, is priviness more likely to decrease than increase?

Decreasing vs Increasing

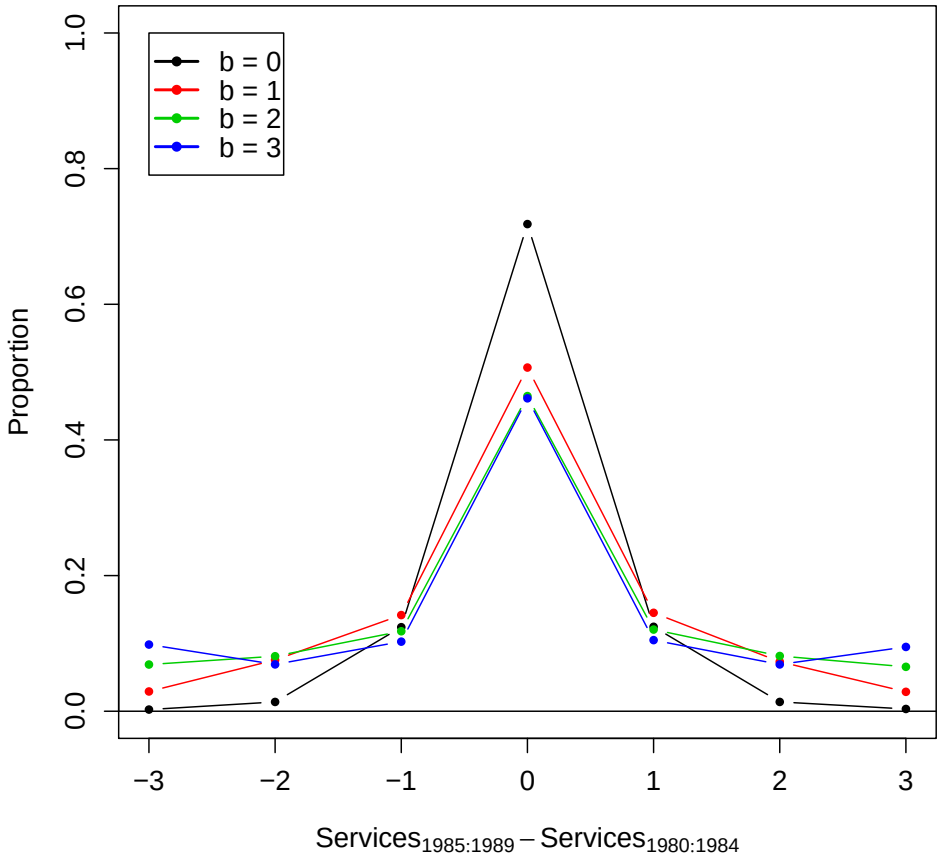


Comparing priviness 1980–1984 against 1985–1989.

Jointness: Do Services Collaborate

- President Reagan decreed that the services (Army, Navy, Air Force) must work together.
- Does the DTIC data support that authors collaborated with those from other services?

Number of Services



Discussion

- Privy analysis provides a quantitative method for horizon scanning.
- Keywords do not seem to be a good surrogate for topic. They do seem to show a drift – do more specific keywords take the place of more general ones?
- Some evidence that authors are more likely to decrease their priviness than increase.
  NB: Recall that this is computed in disjoint windows.
- Large changes in priviness are relatively rare.
- Challenges exist. Much to be done.

Questions?

Contact: david.marchette@navy.mil, dmarchette@gmail.com