Quantitative Horizon Scanning for Mitigating Technological Surprise: Detecting the Potential for Collaboration at the Interface[†]

Carey E. Priebe[1]*, Jeffrey L. Solka[2], David J. Marchette[2] and Avory C. Bryant[2]

[1] *Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218-2682, USA*

[2] *Naval Surface Warfare Center, Dahlgren, VA, USA*

Received 14 September 2010; revised 17 January 2012; accepted 11 February 2012
DOI:10.1002/sam.11143
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: 'The identification of potential breakthroughs before they happen' is a vague data analysis problem and 'the scientific literature' is a massive, complex dataset. Hence QHS for MTS might seem to be prototypical of the data miner's lament: 'Here's some data we have... can you find something interesting?' Nonetheless, the problem is real and important, and we develop an innovative statistical approach thereto—not a final etched-in-stone approach, but perhaps the first complete quantitative methodology explicitly addressing QHS for MTS. © 2012 Wiley Periodicals, Inc. Statistical Analysis and Data Mining, 2012

Keywords: statistical inference; text processing; graphs and networks

1. INTRODUCTION

> 'Future breakthroughs will stem from the fusion of knowledge from different fields ... '

The above sentiment has been expressed many times by many people in recent times [1–4]. Signals extracted from the scientific literature concerning emerging technologies may be useful in predicting when and where technological breakthroughs might be expected [2,3,5–14]. These two ideas together suggest our main premise: that emerging relationships between *disparate fields* may presage potential scientific breakthrough, and that prediction of such breakthroughs based on analysis of the scientific literature may be possible. In his 1976 paper 'Guarding Against Technological Surprise', George Heilmeier wrote that '[t]echnological surprise is not a term that conforms to but one definition' [15]. For our purposes, 'Mitigation of Technological Surprise' (MTS) is defined here as the identification of potential breakthroughs before they happen [16]. (Note that it is the *surprise* we wish to mitigate; not the *technology*.) In this context, 'Horizon Scanning' (analogous to early warning radar) is defined here as the process of systematically exploring the external environment (in this case, the scientific literature) in an effort to detect emerging trends [17]; 'Quantitative Horizon Scanning' (QHS) refers to the use of statistical inference methodologies to this end, as opposed to having human analysts perusing the literature for signs of potential breakthrough. This latter approach—relying solely on human analysts' perusal of the literature—may work well when breakthroughs come from within individual scientific fields for which dedicated subject matter experts are available; alas, it is no longer the case that we have available to us human analysts with expertise across the comprehensive range of scientific endeavor [18]. In mathematics alone, Jacques Hadamard (1865–1963) has been described as 'one of the last universal mathematicians,' [19] and modern mathematics, science and technology as a whole is even less amenable to universalists. Thus, in this article, we provide a quantitative methodology based on analysis of the scientific literature for automatically identifying emerging relationships between *disparate fields* that may presage potential scientific breakthrough. In the grand tradition of statistical prediction, our goal here is to *reduce the search space* of potential emerging relationships between disparate fields.

Unfortunately, 'the identification of potential breakthroughs before they happen' is a vague data analysis

Correspondence to: Carey E. Priebe (cep@jhu.edu).

[†] This article is based on a Keynote Address given by one author (C.E.P.) at QMDNS 2010, May 25–26, Fairfax, VA, USA (presentation slides available at http://www.ams.jhu.edu/∼priebe/qhs4mts.html).

problem and 'the scientific literature' is a massive, complex dataset. Hence QHS for MTS might seem to be prototypical of the data miner's lament: 'Here's some data we have... can you find something interesting?' [20]. Nonetheless, the problem is real and important, and we attempt in this manuscript to develop an innovative statistical approach thereto—not a final etched-in-stone approach, but perhaps the first end-to-end quantitative methodology explicitly addressing QHS for MTS.

2. QHS FOR MTS: PRINCIPLES, OBJECTIVE AND INFERENCE TASK

Our QHS for MTS guiding principles are (1) technological surprise often involves the (unanticipated) fusion of ideas from disparate subject areas, and (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. On the basis of these principles, we develop an inferential approach based on the analysis of scientific literature which can be used as a predictor of the potential for technological surprise.

On the basis of the foregoing principles and the formulation we develop in the sequel, we propose a well-defined QHS for MTS objective: We wish to identify disparate scientific subjects across which a target author set has differential collaborative potential with respect to a baseline author set.

For illustration, we will present a concrete special case of QHS for MTS inference; this example fits within a more general QHS for MTS framework. Let $\mathcal{Y}_1$ and $\mathcal{Y}_2$ denote two collections of authors, and let ℓ_1 and ℓ_2 denote two technology subject categories. We develop a statistic which captures information concerning 'priviness'—the degree of sharing in secret or private knowledge—of author sets to technology categories. This statistic allows for a well-defined QHS for MTS inference task: Large values of the statistic indicate that author set $\mathcal{Y}_2$ is at risk of being technologically surprised by author set $\mathcal{Y}_1$ at the (ℓ_1, ℓ_2) technology interface. Ergo, finding such pairs is a QHS for MTS inference task. Given author sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$, our approach suggests that further investigation is warranted when (1) the statistic is operationally and statistically significant, and (2) some measure $s(\ell_1, \ell_2)$ of the surprise factor for technology category pair (ℓ_1, ℓ_2) is large.

To preview our example presented in Section 5 below, consider a collection of computer science journal articles. The author sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ correspond to authors identified with Great Britain and Germany, respectively. The articles are clustered into technology subject categories $\ell_1 \cdots \ell_m$, and for each (ℓ_i, ℓ_j) pair a log-odds-ratio statistic is used to assess differential collaborative potential for the two author sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ at the (ℓ_1, ℓ_2) technology interface.

3. DATA PROCESSING

Statistical analysis of scenarios comprising multiple modes of association among a collection of actors over time is of ever-increasing importance in a wide-ranging array of applications; for example, communications analysis—who talks to whom, about what, and when. Random graph models are commonly used to model association among actors, and time series of attributed multigraphs, wherein vertex attributes encode relevant information about the individual actors and edge attributes encode modes of association between actors, are a natural extension. Our time series of attributed multigraphs is derived from (1) a collection $M = \{(\mathcal{A}_i, t_i, x_i)\}$ of $(participants, time, content)$ scientific collaboration events (documents) where $\mathcal{A}_i$ is a subset of the collection of all actors $\mathcal{A}$ (e.g., $\mathcal{A}_i$ denotes the authors of the document), time $t_i \in \mathbb{R}_+$ (e.g., the publication date of the document), and x_i is the content of the collaboration (e.g., the content of the document), and (2) attribution functions for vertices and edges.

Let $\mathcal{X}$ be a finite set, representing a collection of text documents from the scientific literature. Consider a collection of functions on $\mathcal{X}$. For our application, these represent *metadata extraction functions* providing *document attributes*. In particular, we have f_1 that extracts document authors, f_2 that extracts document country affiliations, f_3 that extracts document institution affiliations, f_4 that extracts document subject identifications, f_5 that extracts document keywords, f_6 that extracts document abstract, ..., and f_K that extracts document time stamp. Thus each $x \in \mathcal{X}$ has associated with it an author collection $f_1(x)$, a country affiliation collection $f_2(x)$, an institution affiliation collection $f_3(x)$,..., and a time stamp $f_K(x)$. Let $\mathcal{A} = \cup_x \; f_1(x)$ be the overall collection of authors active in $\mathcal{X}$. Let $\mathcal{C} = \cup_x \; f_2(x)$ be the overall collection of countries active in $\mathcal{X}$. Let $\mathcal{I} = \cup_x \; f_3(x)$ be the overall collection of institutions active in $\mathcal{X}$. From this, we can associate *author attributes* with each $a \in \mathcal{A}$, such as country affiliation $g_2(a) \subset \mathcal{C}$, institution affiliation $g_3(a) \subset \mathcal{I}$, etc.

Let $G = (\mathcal{A}, E)$ be our social network on $\mathcal{A}$, where E is the set of edges. For instance, we have available $G = (\mathcal{A}, E; \mathcal{X})$, the coauthorship graph induced by $\mathcal{X}$, where $uv \in E \iff \exists x \in \mathcal{X} \; s.t. \; \{u, v\} \subset f_1(x)$. (Here we are ignoring the *hyper* aspect of edges, for simplicity.) Define $N_k[a]$ to be the collection of vertices v such that there exists a path in G of length less than or equal to k connecting a to v. We can augment G with edges based on institution affiliation, etc. We can also incorporate information external to $\mathcal{X}$, such as graduate student/postdoc connections, etc.

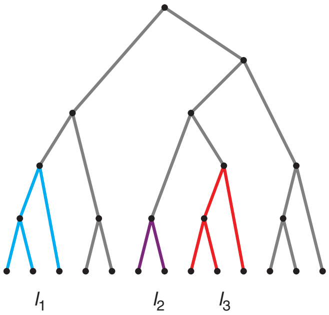

Fig. 1 Notional depiction of subjects of interest derived from a document cluster tree. Any node in the tree might identify a subject of interest ℓ, and then all documents (leaves) under that node are identified with subject ℓ. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Let $i_1, \ldots, i_c$ identify author attributes and let $y_i \subset \cup_{a \in \mathcal{A}} g_i(a)$ denote a specified set of target values. Then $\mathcal{Y} = \{a : g_i(a) \subset y_i \;\forall\; i \in \{i_1, \ldots, i_c\}\}$ is a collection of authors of interest. In particular, with $i = 2$ and $y_i = \{y^*\}$ with $y^* \in \mathcal{C}$, $\mathcal{Y}$ represents authors associated with a country of interest.

Abstractly, we will let any subset $\ell \subset \mathcal{X}$ represent a subject—the subject implicitly defined by the collection of documents ℓ. For example, if we let $H(\mathcal{X})$ be a clustering tree of $\mathcal{X}$, then any node in the tree might identify a subject of interest (see Fig. 1).

Given $\ell \subset \mathcal{X}$, let G_ℓ be the social network $G = (\mathcal{A}, E)$ where edges $uv \in E$ are ℓ-colored based on coauthored documents in ℓ. That is, $uv \in E$ is ℓ-colored $\iff \exists x \in \ell \; s.t. \; \{u, v\} \subset f_1(x)$. For $a \in \mathcal{A}$, let $d_\ell(a)$ denote graph distance in G_ℓ from vertex a to an ℓ-colored edge;

$$d_\ell(a) = \min\{k : \text{ there is an } \ell\text{-colored edge incident}$$
$$\text{to a vertex } v \in N_k[a]\}.$$

For example, if there is an ℓ-colored edge incident to a, then $d_\ell(a) = 0$; if $d_\ell(a) > 0$ and there is an ℓ-colored edge incident to a vertex $v \in N_1[a]$, then $d_\ell(a) = 1$, etc.

4. DETECTION STATISTIC

4.1. Privy Sets

Given $b \in \mathbb{Z}_+$, author subset $\mathcal{Y} \subset \mathcal{A}$, and document subset $\ell \subset \mathcal{X}$, 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 (see Fig. 2).

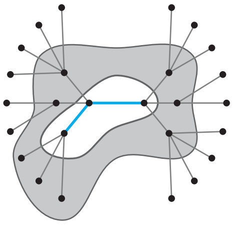

Fig. 2 Notional depiction of the privy property. The three inner-most vertices are 0-privy to the topic associated with the two edges in the inner-most white region, the vertices in the annulus are 1-privy, and the outer vertices are 2-privy. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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

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

(Note that $p^{\mathcal{Y}}$ depends on $b, \ell_1, \ell_2, \mathcal{X}'$ 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 $\mathcal{X}'$) that $\mathcal{Y}$-authors are b-privy to the (ℓ_1, ℓ_2)-interface.

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

DEFINITION: The statistic

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

is the log-odds-ratio for differential b-collaborative potential at the (ℓ_1, ℓ_2)-interface for $\mathcal{Y}_1$ with respect to $\mathcal{Y}_2$. (Any of the various smoothing approaches can be employed to address the case of 0s in the 2 × 2 tables.)

Let ℓ_i and ℓ_j be leaves in $H(\mathcal{X})$. Let $s(\ell_i, \ell_j)$ be some distance (e.g., graph distance in $H(\mathcal{X})$, or embedding distance, or an *a priori* distance provided by expert knowledge) between ℓ_i and ℓ_j; the larger the value of $s(\ell_i, \ell_j)$, the more ‘surprising’ is a collaboration which spans ℓ_i and ℓ_j.

Based on the foregoing formulation, we reiterate our QHS for MTS objective: We wish to identify disparate subjects across which a target author set has differential collaborative potential with respect to a baseline author set.

Statistically significant large values of $\widehat{L}(b, \mathcal{Y}_1, \mathcal{Y}_2, \ell_i, \ell_j,$ $\mathcal{X})$ for which $s(\ell_i, \ell_j)$ is large indicate that $\mathcal{Y}_2$ is at risk of being technologically surprised by $\mathcal{Y}_1$ at the (ℓ_i, ℓ_j)-interface. (Note that p-values are available for $H_0 : L \leq 0$ vs. $H_A : L > 0$ for appropriate sampling distributions on the attributed graphs.) Ergo, finding such pairs is a QHS inference task for MTS. Given one-sided p-values

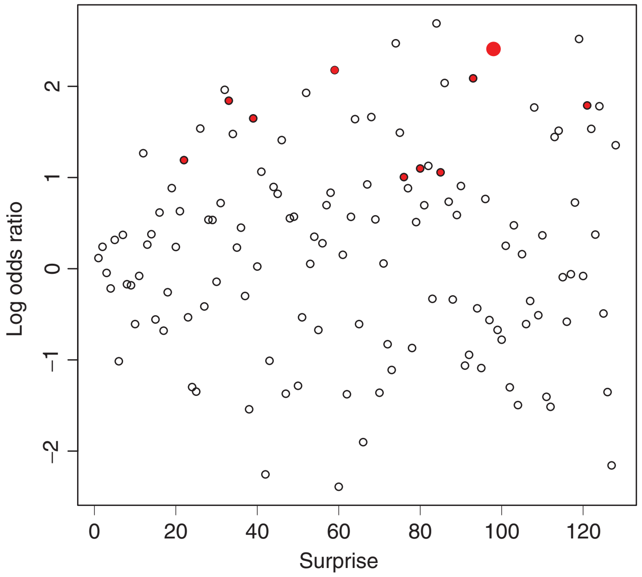

Fig. 3 Notional depiction of $\widehat{L}_{ij}$ vs. s_{ij}. Each dot represents a subject pair, with statistical significance ($p \le \tau$) indicated in red. The large red dot is the first choice for further investigation—the subject pair maximizing some user-specified function $f(\widehat{L}, s, 1/p)$ which is monotonically increasing in all three arguments. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

p_{ij} for $\widehat{L}(b, \mathcal{Y}_1, \mathcal{Y}_2, \ell_i, \ell_j, \mathcal{X})$ for the collection of pairs (ℓ_i, ℓ_j) under consideration, our approach suggests that pairs (ℓ_i, ℓ_j) for which p_{ij} is small and $\widehat{L}_{ij}$ and s_{ij} are large warrant further investigation. See Figure 3. Values in a user-defined region of the upper-right corner of this plot are candidates. (A correction for multiple comparisons can be employed; however, since the p-values are used here simply for selection of pairs of interest, such a correction may be unnecessary in many applications.)

4.2. Incorporating Time

Let $I \subset \mathbb{R}_+$ be a time interval, and consider $\mathcal{X}_I = \{x \in \mathcal{X}\ s.t.\ f_K(x) \in I\}$ to be the documents with time stamp in I. Given $\ell \subset \mathcal{X}$, let $\ell_I = \ell \cap \mathcal{X}_I$, and let $d_{\ell_I}(a)$ denote graph distance in G_{ℓ_I}. Then the time-dependent b-privy sets are defined by

$$A(b, \mathcal{Y}, \ell_I) = \{a \in \mathcal{Y}\ s.t.\ d\ell_I(a) \le b\}.$$

That is, $A(b, \mathcal{Y}, \ell_I)$ denotes all authors in $\mathcal{Y}$ who were b-privy to subject ℓ during time interval I. Given time intervals $I_1, I_2 \subset \mathbb{R}_+$, and altering our log-odds-ratio statistic to consider just a single author set but two time intervals, $\widehat{L}_{I_1,I_2}(b, \mathcal{Y}, \ell_1, \ell_2, \mathcal{X})$ provides a statistic for detecting differential b-collaborative potential at the (ℓ_1, ℓ_2)-interface for $\mathcal{Y}$ over time.

The Privy Prediction Conjecture: Let $Q_I(b, \mathcal{Y})$ be the collection of $\{\ell_1, \ell_2\}$ pairs such that $|A(b, \mathcal{Y}, \ell_{1_I}) \cap A(b, \mathcal{Y}, \ell_{2_I})| > 0$. Let $b' > b$. Consider the conditional probability conjecture that

$$\begin{aligned} &P[\{\ell_1, \ell_2\} \in Q_{I_2}(b, \mathcal{Y}) \mid \{\ell_1, \ell_2\} \notin Q_{I_1}(b, \mathcal{Y})\ \& \\ &\quad \{\ell_1, \ell_2\} \in Q_{I_1}(b', \mathcal{Y})] > \\ &P[\{\ell_1, \ell_2\} \in Q_{I_2}(b, \mathcal{Y}) \mid \{\ell_1, \ell_2\} \notin Q_{I_1}(b, \mathcal{Y})]. \end{aligned}$$

For example, with $b = 0$ and $b' = 1$, we are interested in the conditional probability that $\mathcal{Y}$ is now 0-privy given that $\mathcal{Y}$ was not previously 0-privy. To the extent that $\mathcal{Y}$ having previously been 1-privy increases the probability that $\mathcal{Y}$ is now 0-privy, identifying 1-priviness (gaining potential) can predict 0-priviness (having potential).

The Privy Prediction Conjecture suggests that privy analysis can hope to predict future breakthroughs stemming from the fusion of knowledge from disparate fields.

5. EXAMPLE FROM THE SCIENTIFIC LITERATURE

We present an example of privy analysis of scientific literature using the Scopus database.

'Scopus is a database of abstracts and citations for scholarly journal articles. It covers nearly 18,000 titles from more than 5,000 international publishers, including coverage of 16,500 peer-reviewed journals in the scientific, technical, medical, and social sciences (including arts and humanities) fields. It is owned by Elsevier…' [21].

We consider a total of $|\mathcal{X}| = 230931$ Computer Science documents from time intervals $I_1 = 1995–1998$ and $I_2 = 1999–2000$. We consider author subsets from all of Scopus identified with Great Britain (gb) and Germany (de); the number of authors is roughly the same for each country and each time interval: $n_{gb,1995-1998} = 218848$; $n_{de,1995-1998} = 229139$; $n_{gb,1999-2000} = 179247$; $n_{de,1999-2000} = 185161$.

The abstracts are clustered into 256 clusters via CLUTO, a freely available software toolkit for clustering [22]. Figure 4 presents an MDS embedding [23] consisting of 48 super-clusters. This clustering and super-clustering will be used to identify *disparate* subjects—two clusters which fall into different super-clusters will be judged sufficiently disparate. (NB: this restricted definition of surprise—as opposed to the purely distance-based definition $s(\ell_i, \ell_j)$ provided above—is suggested by the requirement for inference on *disparate fields across which no single human analyst has sufficient expertise*. Indeed, if documents are clustered explicitly by available subject matter expertise, providing super-clusters as in Figure 4, then this restricted definition of surprise corresponds directly with

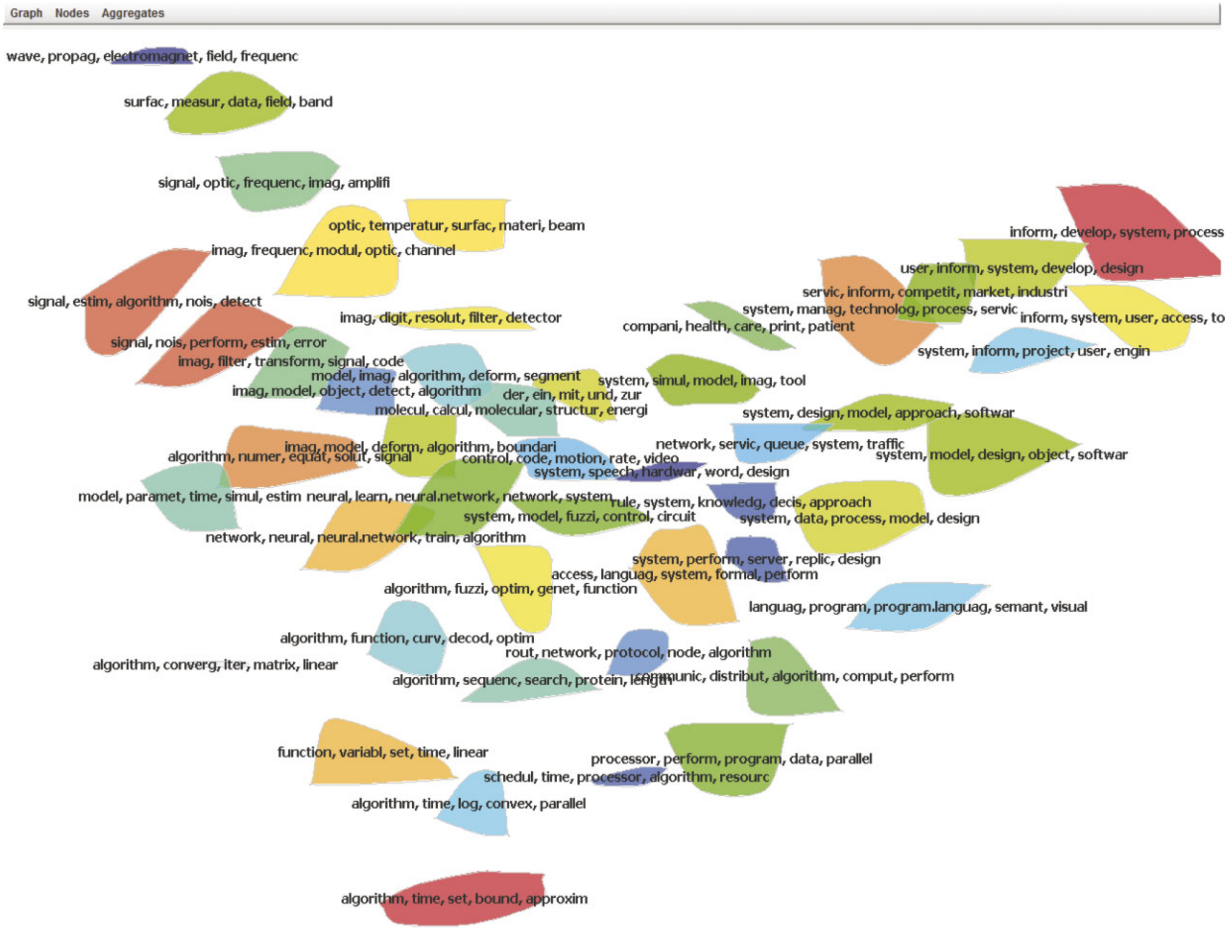

Fig. 4 MDS embedding of 48 super-clusters of 230931 Computer Science abstracts from 1995 to 2000. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the ‘disparate subject areas’ requirement of our QHS for MTS exploitation task.)

Figure 5 presents results from our privy analysis, with one dot for each of the $\binom{256}{2}$ cluster pairs. The left plot is for $b = 1$ at I_1: red denotes 1-privy, not 0-privy, and statistically significant (here we consider the two-sided test) at I_1; black denotes 1-privy, not 0-privy, and *not* statistically significant at I_1; and green denotes also 0-privy at I_1. Thus the red and black together provide the candidate cluster pairs for moving to 0-priviness (having potential), and the red provide the candidate cluster pairs which are already 1-privy (gaining potential). The right plot is for $b = 0$ at I_2; color is inherited from I_1. There are 29 big filled dots denoting statistically significant cluster pairs for $b = 0$ at I_2: 7 green, 6 black, and 16 red. From the Privy Prediction Conjecture, we define the Q_{factor} to be the ratio

$$Q_{factor} = \frac{\substack{P[\{\ell_1,\ell_2\} \in Q_{I_2}(b,\mathcal{Y}) \mid \{\ell_1,\ell_2\} \notin Q_{I_1}(b,\mathcal{Y}) \\ \&\{\ell_1,\ell_2\} \in Q_{I_1}(b',\mathcal{Y})]}}{P[\{\ell_1,\ell_2\} \in Q_{I_2}(b,\mathcal{Y}) \mid \{\ell_1,\ell_2\} \notin Q_{I_1}(b,\mathcal{Y})]}.$$

If this ratio is greater than one, then identifying 1-priviness (gaining potential) can predict 0-priviness (having potential). For this data, we obtain $Q_{\text{factor}} \approx 2.5$. The maximum surprise statistically significant red dot in the right-hand plot ($\arg\max_{\text{topicpair}(i,j)} f(|\widehat{L}_{ij}|, s_{ij}, 1/p_{ij})$) denotes two clusters which are in distinct super-clusters. Therefore, privy analysis can hope to predict future breakthroughs stemming from the fusion of knowledge from disparate fields. (Human retrospective analysis is necessary to determine which, if any, of the indicated fusion candidates are of actual practical concern.)

6. EXAMPLE MATHEMATICAL MODEL

> ‘That’s all well & good in *practice*, but how does it work in *theory*?’

We present here a sketch outline for a mathematical model relevant to our privy analysis framework and to the example from the scientific literature presented above. The

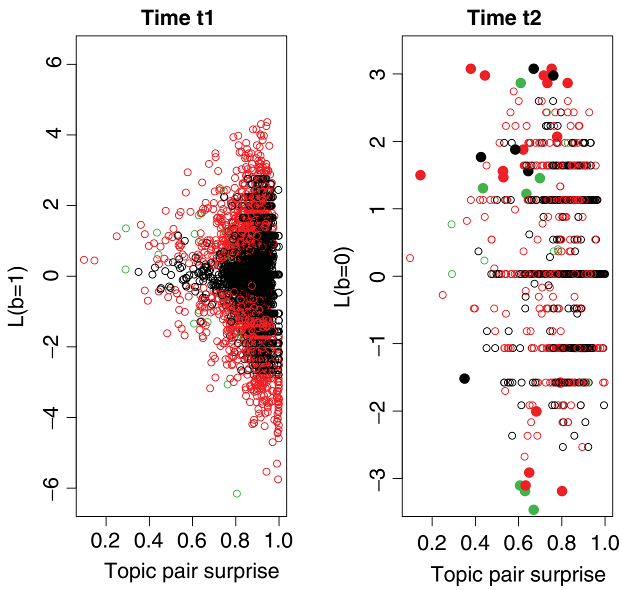

Fig. 5 Results from privy analysis of Computer Science data set. See text for description. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

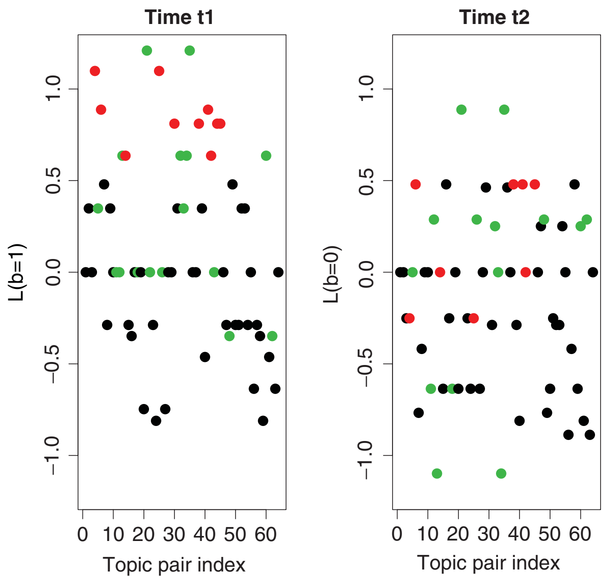

Fig. 6 Simulation results indicating privy prediction capabilities. See text for description. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

purpose of this section is to provide hints as to useful theoretical constructs in this context.

Lee and Priebe [24] present a latent process model for time series of attributed graphs. For each $v \in V$, let A_v be a continuous-time, $(K+1)$-state stochastic process. For each $k \in \{0\} \cup [K]$, let $X_{v,k}(t) = \int_{t-1}^{t} 1_{\{k\}}(A_v(s))ds$; $X_{v,k}(t)$ is the amount of time that the process A_v occupies state k during interval $(t-1, t]$, so that $\sum_{k=0}^{K} X_{v,k}(t) = 1$. Let $p_{uv,k}(t) = X_{u,k}(t)X_{v,k}(t)$. At time $t \in \mathbb{N}$, for each $\{u, v\} \in \binom{V}{2}$, attributed edge $uv(t)$ is discrete random variable on $\{0, 1, \ldots, K\}$ with probability vector $[1 - \sum_{k=1}^{K} p_{uv,k}(t), p_{uv,1}(t), \ldots, p_{uv,K}(t)]^T$, thus defining graph $G(t)$. Now consider '*augmentation from the past*': $\vec{X}'_v(t) = (\vec{X}_v(t) + c_v(t)\vec{\theta}_v(t))/(1 + c_v(t)||\vec{\theta}_v(t)||_1)$ where $c_v(t) \in [0, \infty)$, $\vec{\theta}_v(t) = [0, \theta_{v,1}(t), \ldots, \theta_{v,K}(t)]^T$, and $\theta_{v,k}(t) = \sum_{e:d(v,e;G(t-1))=1} I\{a(e) = k\}/|\{e : d(v, e; G(t-1)) = 1\}|$. Our time series of graphs is produced as above, with bias introduced into the dot product probability computation based on the graph at time $t-1$, using $p_{uv,k}(t) = X'_{u,k}(t)X'_{v,k}(t)$.

Figure 6 presents simulation results from this model, analogous to the experimental results presented above, indicating privy prediction capabilities. The left plot is for $b = 1$ at I_1: red denotes 1-privy, not 0-privy, and large at I_1; black denotes 1-privy, not 0-privy, and *not* large at I_1; and green denotes also 0-privy at I_1. The right plot is for $b = 0$ at I_2; color is inherited from I_1. The plot indicates behavior relevant to privy prediction which holds by model construction: red are stochastically larger than black.

THEOREM: Let $c_v(t) > 0$. Then the Privy Prediction Conjecture holds. Sketch:

$$\begin{aligned}
&P[v \in A(0, \mathcal{Y}, \ell_t)|v \notin A(0, \mathcal{Y}, \ell_{t-1})] \\
&\quad = wP[v \in A(0, \mathcal{Y}, \ell_t)|v \notin A(0, \mathcal{Y}, \ell_{t-1}) \text{ and} \\
&\quad v \in A(1, \mathcal{Y}, \ell_{t-1})] + (1-w)P[v \in A(0, \mathcal{Y}, \ell_t)|v \\
&\quad \notin A(0, \mathcal{Y}, \ell_{t-1}) \text{ and } v \notin A(1, \mathcal{Y}, \ell_{t-1})] \\
&\quad < P[v \in A(0, \mathcal{Y}, \ell_t)|v \notin A(0, \mathcal{Y}, \ell_{t-1}) \text{ and} \\
&\quad v \in A(1, \mathcal{Y}, \ell_{t-1})]
\end{aligned}$$

by construction.

Now pass from author sets A to subject pair sets Q.

7. CONCLUSIONS

We have produced theory and methods for Quantitative Horizon Scanning (QHS) for Mitigation of Technological Surprise (MTS), approaching the task from the standpoint of mathematical statistics—with explicit guiding principles and well-defined inferential goals. The illustrative example presented above provides a concrete special case within a more general QHS for MTS framework which we have elucidated. The result is perhaps the first true inferential theory for QHS for MTS.

In summary, we have presented motivation, a quantitative problem formulation, and a statistical inference methodology for QHS for MTS. Our *guiding principles* are (i)

Technological surprise often involves the (unanticipated) fusion of ideas from disparate subject areas, and (ii) 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. Our *objective* is to identify disparate subjects across which a target author set has differential collaborative potential with respect to a baseline author set. Our *inference task* is to find document subset pairs for which $|\widehat{L}(b, \mathcal{Y}_1, \mathcal{Y}_2, \ell_i, \ell_j, \mathcal{X})|$ is statistically significantly large.

Our solution formulation has admittedly given short shrift to numerous important practical issues, and sticky wickets abound. We discuss here a few of the most pressing.

(i) Text document processing and clustering is a rich and active field with many and various competing methodologies (see, e.g., [25,26] and the myriad references provided therein); the methodologies we have employed are simple, off-the-shelf, for illustrative purposes. Investigation of the robustness of privy analysis to these choices is ongoing. (We conjecture that privy analysis is more sensitive to graph construction details and author ambiguities than to text document processing and clustering methodologies.)

(ii) Author disambiguation (see [27] for a recent review) is an important aspect of the type of analysis we have outlined; our approach will suffer significantly if this disambiguation is poorly done. (In particular, while SCOPUS has unique author identifications which in theory provide a solution to the author disambiguation problem, Science Citation Index has no such built-in solution.) We recommend theoretical, simulation, and experimental investigation of the effect of author ambiguity in the context of our privy analysis.

(iii) The social network construction is of paramount importance to our privy analysis. In fact, we suggest that our approach will be robust to document processing and clustering vagaries but sensitive to small changes in the graph structure. Elaborate Knowledge Base-induced social structures [11,28–31] might be relevant, and various methodological alterations to our approach—e.g. expected commute time or diffusion distance as opposed to shortest path distance—suggest themselves. We are considering the cardinality of author sets only, but not all vertices are created equally—some authors are more influential than others and should carry more weight in our privy analysis. Similarly, not all edges are created equally—some papers are more important than others, and perhaps a maturity index such as the Technology Readiness Level [32] could be usefully employed. We have considered priviness as a binary property. Yet someone who has written multiple articles on a topic may be considered to be more knowledgeable than someone who has written just one, suggesting that a weighted privy analysis is appropriate. Also, we have considered shortest-path priviness. As mentioned above, some measure of diffusion or expected commute time could be used instead. This would utilize the number of paths to a topic rather than simply the shortest, and could incorporate weights according to the number of publications and other measures of importance of the papers and the authors. Furthermore, priviness need not be defined solely through co-authorship. For example, it is clear that if a paper cites another paper on a topic, this is evidence that the authors are privy to that topic. The citation graph can be used as a second graph (either on the same vertices (authors) as the co-authorship graph, or on the edges (papers)). Utilizing both types of information should produce superior inference. There are folk theorems that state that everyone is no more than six steps away from everyone else (seemingly no matter how the social network is defined). Without delving into the details of such claims, we note that there is some sense in which larger values of b become less and less meaningful in our privy analysis. A diffusion approach would down-weight longer paths, and perhaps there is a natural cut-off ($b = 2$?) beyond which topic information is unlikely to flow. Such ideas can be incorporated in the kernel of the diffusion operator or as simply a hard threshold on calculations.

(iv) Time introduces several important considerations. The first is the question of how to quantify memory/currency: should time affect how strongly one is b-privy today, given that the last time one was b-privy was t years ago? The issues of memory (is it likely that one still remembers one's older work) and currency (is the work still relevant to the topic today) need to be addressed. This is related to a second issue: topics change in time. This is more than the fact that science progresses and hence the work of the past becomes less current than the work of the present, but also the fundamental science of a topic can change. For example, molecular chemistry has wrought fundamental changes in biology, and the language used has changed enough that the distance between the topics 'biology' in 1970 and 'biology' in 2010 may be as large as that between 'biology' and 'physics' in 1970. In addition, new topics appear and old topics disappear as science progresses, and this is an important consideration in QHS.

(v) Regarding the inference itself—the odds-ratio p-value—we note that our odds-ratio is *formally* the same as that for which inference is available, but sampling details are relevant. There is also the issue of multiple comparisons affecting any omnibus inference, which we have ignored for the nonce. (For the exploitation task considered herein—individual topic-pair inferences for the purpose of identifying and ranking candidates—we need not grapple with the multiple comparison issues.)

(vi) Finally, we note the need for, and difficulty of, human retrospective analysis for evaluating methodologies for QHS for MTS. This issue seems to us to be of critical importance, and we recommend a large-scale effort toward this end. Evaluation of the detection performance, sensitivity, and specificity of a method for QHS for MTS is a difficult problem. One can do retrospective studies, but even here it is difficult to decide on objective criteria for 'surprise' and 'breakthrough', beyond the basic ones we have articulated here. At a minimum, one would want to look at the future papers by the authors in question and determine if they are indeed 'at the interface.' One could also measure whether the two topics are moving closer together, indicating a fusing of the two topic areas, or if a new topic is forming that incorporates information from both topics. One could bring in other data, such as patents, news reports, or other sources of information about the purported breakthrough. It is a challenging problem to design such experiments so that unbiased evaluations are possible.

Despite these myriad issues, we have demonstrated, through theory, simulation, and experiment, that identifying 1-priviness (gaining potential) predicts 0-priviness (having potential)—$Q_{\text{factor}} \gg 1$. Thus we have shown that privy analysis can hope to predict future breakthroughs stemming from the fusion of knowledge from disparate fields.

ACKNOWLEDGMENTS

This work is partially supported by National Security Science and Engineering Faculty Fellowship (NSSEFF), Air Force Office of Scientific Research (AFOSR), Office of Naval Research (ONR), Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), the American Society for Engineering Education (ASEE) Sabbatical Leave Program, and the U.S. Navy In-house Laboratory Independent Research (ILIR) Program.

REFERENCES

[1] P. Beckman, D.o.e. supercomputing resources available for advancing scientific breakthroughs, 2009. http://www.azom.com/news.asp?newsID=16553.

[2] R. K. Buter, E. C. M. Nyons, and A. F. J. V. Raan, Searching for converging research using field to field citations, Scientometrics (2010). DOI: 10.1007/s11192-010-0246-0.

[3] L. Fleming and O. Sorenson, Technology as a complex adaptive system: evidence from patent data, Res Policy 30(7) (2001), 1019–1039.

[4] L. E. Johnston, Language, culture, and cooperation in scientific and technical intelligence, 2007. Center for the Study of Intelligence. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol52no2/meeting-a-critical-challenge.html.

[5] Q. Mei and C. Zhai, Discovering evolutionary theme patterns from text: an exploration of temporal text mining, KDD'05: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM, 2005, 198–207.

[6] G. Silverberg and B. Verspagen, A percolation model of innovation in complex technology spaces, J Econ Dyn Cont 29(1–2) (2005) 225–244. http://ideas.repec.org/a/eee/dyncon/v29y2005i1-2p225-244.html.

[7] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and C. L. Giles, Detecting topic evolution in scientific literature: how can citations help? Proceeding of the 18th ACM Conference on Information and Knowledge Management, 2009, ACM, 957–966, http://dblp.uni-trier.de/db/conf/cikm/cikm2009.html#HeCPQMG09.

[8] W. Peng and T. Li, Author-topic evolution analysis using three-way non-negative paratucker, SIGIR'08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2008, 819–820.

[9] T. L. Griffiths and M. Steyvers, Finding scientific topics, Proc Natl Acad Sci 101(Suppl. 1) (2004), 5228–5235.

[10] C. Chen, Y. Chen, M. Horowitz, H. Hou, Z. Liu, and D. Pellegrino, Towards an explanatory and computational theory of scientific discovery, J Inform 3(3) (2009), 191–209.

[11] J. Wang, X. Geng, K. Gao, and L. Li, Study on topic evolution based on text mining, Fourth International Conference on Fuzzy Systems and Knowledge Discovery 2 (2008), 509–513.

[12] M.-J. Shih, D.-R. Liu, and M.-L. Hsu, Mining changes in patent trends for competitive intelligence, PAKDD'08: Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, vol. 5012, Springer-Verlag, Lecture Notes in Computer Science, 2008, 999–1005.

[13] T. U. Daim, G. Rueda, H. Martin, and P. Gerdsri, Forecasting emerging technologies: Use of bibliometrics and patent analysis, Technological Forecasting and Social Change 73(8) (2006), 981–1012. http://dx.doi.org/10.1016/j.techfore.2006.04.004.

[14] R. Ohniwa, A. Hibino, and K. Takeyasu, Trends in research foci in life science fields over the last 30 years monitored by emerging topics, Scientometrics, June 2010, http://dx.doi.org/10.1007/s11192-010-0252-2.

[15] G. Heilmeier, Guarding against technological surprise, Air University Review, 1976.

[16] Avoiding technology surprise for tomorrow's warfighter: A Symposium Report. Committee for the Symposium on Avoiding Technology Surprise for Tomorrow's Warfighter; National Research Council, Standing Committee on Technology Insight-Gauge, Evaluate & Review (TIGER), ISBN-10: 0-309-14228-8 2009.

[17] United Kingdom Chief Scientist Advisors Committee 2004, 2004.

[18] A. Robinson, *The Last Man Who Knew Everything: Thomas Young, The Anonymous Polymath Who Proved Newton Wrong, Explained How We See, Cured the Sick, and Deciphered the Rosetta Stone, Among Other Feats of Genius*, Pi Press, 2005.

[19] V. G. Mazia and T. O. Shaposhnikova, Jacques Hadamard: A Universal Mathematician (History of Mathematics, Vol. 14), American Mathematical Society, 1998.

[20] M. J. A. Berry and G. S. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, (2nd ed), Wiley Computer Publishing, 2004.

[21] Wikipedia. Scopus, 2010, http://en.wikipedia.org/wiki/Scopus (accessed May 1, 2010).

[22] G. Karypis, Cluto-A Clustering Toolkit, Tech. rep., University of Minnesota, 2002.

[23] J. L. Solka and A. C. Bryant, Multi-feature clustering and visualization of large document collections, Stat Anal Data Mining, 2011, in press.

[24] N. H. Lee and C. E. Priebe, A latent process model for time series of attributed random graphs, Stat Infer Stoch Proc, 14(3) (2011), 231–253.

[25] M. W. Berry, Survey of Text Mining I: Clustering, Classification, and Retrieval (No. 1), Springer, 2003.

[26] M. W. Berry, Survey of Text Mining II: Clustering, Classification, and Retrieval (No. 2), Springer, 2007.

[27] N. R. Smalheiser and V. I. Torvik, Author name disambiguation, Annu Rev Inform Sci Technol (ARIST) 43 (2009).

[28] S. Asur, S. Parthasarathy, and D. Ucar, An event-based framework for characterizing the evolutionary behavior of interaction graphs, ACM Trans Knowl Discov Data 3(4) (2009), 1–36.

[29] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, Group formation in large social networks: membership, growth, and evolution, KDD ’06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2006, 44–54. http://dx.doi.org/10.1145/1150402.1150412.

[30] L. M. Bettencourt, D. I. Kaise, and J. Kaur, Scientific discovery and topological transitions in collaboration networks, J Inform 3(3) (2009), 210–221. http://www.sciencedirect.com/science/article/B83WV-4W8TJP2-5/2/dc9439b7619e0c9483a8378e99722e30. Science of Science: Conceptualizations and Models of Science.

[31] M. Lin and N. Li, Scale-free network provides an optimal pattern for knowledge transfer, Phys A: Stat Mech Appl 389(3) (2010), 473–480. http://www.sciencedirect.com/science/article/B6TVG-4XFGJC3-2/2/2684720d6f4057eaa30c69a27c578b5b.

[32] D. A. Sparrow and S. Cazares, How dod’s tra process could be applied to intelligent systems development?, PerMIS 2007: Proceedings of the 2007 Workshop on Performance Metrics for Intelligent Systems, ACM, 2007, 35–39.