DESIGNING DISCOVERY
Hypergraph Models of Innovation for Science & Technology

James Evans
Feng (Bill) Shi
Big Data, Machine Learning and Intelligent Crowdsourcing enables us to:

1. Trace
2. Understand
3. Discover
4. Improve...the scientific and scholarly process

computationally enhanced

Science of Science
How Science Thinks
**Fig. 2.** Stable strategies and their dynamical origin.

(A) The empirical frequency of each strategy (solid line) with 95% confidence intervals smaller than the solid lines (see SOM). Dotted lines show the predictions of our generative model. In 1980 (black arrow) chemical annotation is introduced in MEDLINE; see (13). This distorts parameter estimates until 1987 (parameters for year $t$ are inferred from the six

Year

STRATEGIES OVER TIME

- **Ratio**
- **Year**
- **1980**
- **1985**
- **1990**
- **1995**
- **2000**
- **2005**

- **repeat bridges**
- **repeat consolidations**
- **new bridges**
- **new consolidations**
- **jumps**
High-throughput quasi-replication

- Credible claims: High support and independence, $N = 53$ (1133/18)
- Promising claims: Moderate support and independence, $N = 865$ (5462/169)
- Exploratory claims: Low support, $N = 29296$ (35738/881)
- Disputed claims: $N = 2376$ (2655/4688)
...and in the social sciences

Estimating original models

- Articles typically have
  - A few dependent variables
  - A few independent variables

Approximate Estimation of Original Models

\[
Y_{1,t} = X_{1,t} \beta + \epsilon \\
Y_{2,t} = X_{2,t} \beta + \epsilon \\
\vdots \\
Y_{t,t} = X_{t,t} \beta + \epsilon
\]

\[
X' = \begin{pmatrix}
1 & \cdots & x_{1,1,t}^* & \cdots & x_{1,t,t}^*
\end{pmatrix}
\]

"Race, Sex and Feminist Outlooks" (Ransford and Miller 1983)

- FEHOME_{1974} = X_1 \beta + \epsilon
- FEWORK_{1974} = X_2 \beta + \epsilon
- FEPOSI_{1974} = X_3 \beta + \epsilon
- FEPOS2_{1974} = X_4 \beta + \epsilon
- FEPOS3_{1974} = X_5 \beta + \epsilon

"Confidence in Science: The Gender Gap" (Fox and Firebaugh 1992)

- CONSCI = X_1 \beta + \epsilon
- CONFINAN = X_2 \beta + \epsilon
- CONBUS = X_3 \beta + \epsilon
- CONCLERG = X_4 \beta + \epsilon
- CONEDUC = X_5 \beta + \epsilon
- COMPRESSION = X_6 \beta + \epsilon
- RELIG = X_7 \beta + \epsilon

The Effect of Substituting One Variable: Perturbed minus Original models

\[
R^2 \\
\tilde{R}^2
\]

Significant effects (p < .05)

Central variables (independent and standardised)

Effect size

All variables (standardised)

Effect size

% Change
Active Learning for Intelligent Survey Design

Which place looks safer?
PREDICTING & GENERATING SCIENTIFIC SUCCESS

• Predict **combination of concepts** in future discoveries & inventions

• Predict **level of impact** for future discoveries
Representation

- Collocation Network / Adjacency Matrix
KNOWLEDGE REPRESENTATION

Representation

- Collocation Network / Adjacency Matrix
- Semantic Graph or Hypergraph

Extraction

- Inexpensive

Inference

- Expensive
- Under
- Over
# Knowledge Representation

<table>
<thead>
<tr>
<th>Representation</th>
<th>Extraction</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collocation Network / Adjacency Matrix</td>
<td>Inexpensive</td>
<td>Over</td>
</tr>
<tr>
<td>Collocation Hypergraph / Adjacency Tensor</td>
<td>Inexpensive</td>
<td>Exact</td>
</tr>
<tr>
<td>Semantic Graph or Hypergraph</td>
<td>Expensive</td>
<td>Under</td>
</tr>
</tbody>
</table>
Graph vs. Hypergraph
Matrix vs. Tensor

- Hypergraph CAN ALSO be rendered as a simple, 2-mode matrix where each set connects to each concept, but this removes all informational geometry.

- Hypergraph rendered as a hyper-matrix or tensor retains its geometry, convexity, etc.
MODELING OPPORTUNITY

Automatically generate promising discoveries

(or feed recipes to scientist chefs)
Atypical Combinations and Scientific Impact

Brian Uzzi\textsuperscript{1,2}, Satyam Mukherjee\textsuperscript{1,2}, Michael Stringer\textsuperscript{2,3}, Ben Jones\textsuperscript{1,4,*}
CONCEPTS = CITED JOURNALS

COMBINATIONS = PAIRWISE FREQUENCY DISTRIBUTION

PREDICTION < 10%
OUR PROJECT

CONCEPTS = CONCEPTS

CONTEXTS = JOURNALS

COMBINATIONS = COMPLETE COMBINATION

PREDICTION > 40%
Mixed-Membership, High-Dimensional Block Model

\[ \lambda = (\theta_{11}\theta_{21}\theta_{31} + \theta_{12}\theta_{22}\theta_{32})r_1r_2r_3 \]

Propensity that this combination will turn into a paper:

Number of papers on this combination: \( X \sim Poisson(\lambda) \)
Generative Model for Hypergraph

- For any combination $h$ of nodes
- Calculate propensity $\lambda_h = \sum_k \prod_{i \in h} r_i \theta_{ik}$
- Draw the number of hyperedges of $h$ from $X_h \sim Poisson(\lambda_h)$
- Likelihood to generate the hypergraph

$$p(G|\theta, r) = \prod_{h \in H} p(x_h|\theta, r)$$
Toy Example

\[ p(G|\theta, r) = \prod_{h \in H} p(x_h|\theta, r) \]
Toy Example—Convergence
Toy Example

\[
\begin{align*}
\lambda &= (\theta_{11} \theta_{21} \theta_{31} + \theta_{12} \theta_{22} \theta_{32}) r_1 r_2 r_3 \\
\end{align*}
\]
Evolution of the Network

Hidden Markov Model

\[ G^0 \rightarrow G^1 \rightarrow G^2 \rightarrow G^3 \]

\( G^t \): observed network at time \( t \)
\( \theta^t \): latent positions of the elements at time \( t \)
Complete Model

- Log-likelihood function

\[ l(\theta_1, ..., \theta_T) = \log P(G_1, ..., G_T | \theta_1, ..., \theta_T) \]

\[ = \sum_{t=1}^{T} \left[ \log P(\theta^t | \theta^{t-1}) + \log P(G^t | \theta^t) \right] \]

\[ = \sum_{t=1}^{T} \left[ \sum_i \sum_k (\theta_{ik}^t - \theta_{ik}^{t-1})^2 / 2\sigma^2 + \sum_{h \in G^t} (x_h \log \prod_k \prod_{i \in h} \theta_{ik}^t - \sum_k \prod_{i \in h} \theta_{ik}^t) \right] \]

- Impossible to optimize! Incomputable

\[ 2^N \text{ possible combinations} \]
Maximal Likelihood Estimate

Algorithm

• Generate $t$ from $1,\ldots,T$ uniformly at random

• For $d = 2, \ldots, D$, pick a random set $H_d^t$ of combinations of order $d$ from $G^t$.

• Calculate $S_d^t = \sum_{h \in H_d^t} [x_h \log \sum \prod \theta_{ik}^t - \sum \prod \theta_{ik}^t]$.

• Approximate $\nabla l(\theta)$ by $\nabla (\sum_d S_d^t)$.

• Update $\hat{\theta} = \hat{\theta} + \eta \nabla l(\hat{\theta})$.
Theorem

Let $f(\theta, t) = \sum_{d=2}^{D} S_d^t$ and $t \sim \text{randint}(1, T)$, then

$$E[\nabla f(\theta, t)] = \nabla l(\theta)$$

Corollary

$\hat{\theta}$ will converge to the maximal likelihood estimate.
Datasets

20M PubMed articles (1865 to 2015)

15,000 MeSH term Concepts (e.g., PCR, hypertension, DNA, testosterone)

.5M APS articles (1880-2015)

80,000 PACS code Concepts (e.g., neutron star core, lie algebras, polarization)

1.5M US Patents

45,000 USPC subclasses (e.g., arc lamp, electrolytic condenser, paper, button)
NAP protects hippocampal neurons against multiple toxins

Ilona Zemlyak\textsuperscript{a,b}, Nathan Manley\textsuperscript{b}, Robert Sapolsky\textsuperscript{b}, Ilvana Gozes\textsuperscript{a,*}

\textsuperscript{a}Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Israel
\textsuperscript{b}Department of Biological Sciences, Stanford University, Stanford, USA

\textbf{A R T I C L E   I N F O}

Article history:
Received 1 July 2007
Received in revised form 6 August 2007
Accepted 7 August 2007
Published on line 11 August 2007

Keywords:
NAP
Kainic acid
Oxygen-glucose deprivation
Sodium cyanide
Neuronal death

\textbf{A B S T R A C T}

The femtomolar-acting protective peptide NAP (NAPVSIPQ), derived from activity-dependent neuroprotective protein (ADNP), is broadly neuroprotective in vivo and in vitro in cerebral cortical cultures and a variety of cell lines. In the present study, we have extended previous results and examined the protective potential of NAP in primary rat hippocampal cultures, using microtubule-associated protein 2 (MAP2) as a measure for neuroprotection. Results showed that NAP, at femtomolar concentrations, completely protected against oxygen-glucose deprivation, and cyanide poisoning. Furthermore, NAP partially protected against kainic acid excitotoxicity. In summary, we have significantly expanded previous findings in demonstrating here direct neuroprotective effects for NAP on vital hippocampal neurons that are key participants in cognitive function in vivo.

\textcopyright 2007 Elsevier Inc. All rights reserved.
NAP protects hippocampal neurons against multiple toxins

Ilona Zemlyak, Nathan Manley, Robert Sapolsky, Ilanna Gozes

Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Israel
Department of Biological Sciences, Stanford University, Stanford, USA

ARTICLE INFO

Article history:
Received 1 July 2007
Received in revised form
6 August 2007
Accepted 7 August 2007
Published on line 11 August 2007

Keywords:
NAP
Kainic acid
Oxygen-glucose deprivation
Sodium cyanide
Neuronal death

ABSTRACT

The femtomolar-acting protective peptide NAP (NAPVSIPQ), derived from activity-dependent neuroprotective protein (ADNP), is broadly neuroprotective in vivo and in vitro in cerebral cortical cultures and a variety of cell lines. In the present study, we have extended previous results and examined the protective potential of NAP in primary rat hippocampal cultures, using microtubule-associated protein 2 (MAP2) as a measure for neuroprotection. Results showed that NAP, at femtomolar concentrations, completely protected against oxygen-glucose deprivation, and cyanide poisoning. Furthermore, NAP partially protected against kainic acid excitotoxicity. In summary, we have significantly expanded previous findings in demonstrating here direct neuroprotective effects for NAP on vital hippocampal neurons that are key participants in cognitive function in vivo.
Network Distance

Authors-Authors

Chemicals-Chemicals

Authors-Chemicals

Authors-Methods
Community Structure

- Marijuana Abuse, Phencyclidine Abuse, Alkaloids, Decompression, Substance Abuse Detection
- Mouth, Jaw, Tooth Diseases, Dental Materials, Oral Medicine
- Eye, Ear, Skin Diseases, Body Injuries
- Parasitic Diseases, Helminthiasis
- Marijuana Abuse, Phencyclidine Abuse, Alkaloids, Decompression, Substance Abuse Detection
Predict New Hyperedges

\[ \lambda = (\theta_{11} \theta_{21} \theta_{31} + \theta_{12} \theta_{22} \theta_{32}) r_1 r_2 r_3 \]
Predicting Papers

- Precision: out of all predicted papers, how many actually happen
- Recall: out of all future papers, how many are predicted to happen

AUC > 0.90
Predicting Papers

- Precision: out of all predicted papers, how many actually happen
- Recall: out of all future papers, how many are predicted to happen

AUC > 0.95
Novelty and Impact

Searching Broadly

Not Citing too Broadly
Novelty and Impact

Searching Broadly

Not Citing too Broadly
Content vs. Context

Content Correlates at <.1 with Context
Context does NOT proxy for Content
Context vs. Content

Impact increases with conventionality of journal combinations [Uzzi, et al.]

Impact (weakly) decreases with conventionality of MeSH term combinations.

Same methodology [Uzzi, et al.], but contradictory results?

1. Technically: Not enough MeSH terms per paper to calculate median and tail.
2. Conceptually: Reference list is intended to situate a paper in the literature
3. Mismatch between content and context?
Humble Innovation:
Novel Exploration and Conservative Claims
Audacious Invention:
Novel Exploration and Outsized Claims
Audacious Invention: Novel Exploration and Outsized Claims when examiner citations are removed
Content vs. Context

Content Correlates at $\sim.1$ with Context
Context does NOT proxy for Content
Novelty and Impact

Novel Knowledge from diverse fields unlikely imagined

Novel but obvious

Interdisciplinary but focused
Novelty and Impact

Novel Knowledge from diverse fields unlikely imagined

Interdisciplinary but focused

Novel but obvious
Doubling Sensitivity with High Dimensionality

- 12-13% - pairwise context (Uzzi’s method)
- 25% - high dimensional context (our method)
- 10% - pairwise content (Uzzi’s method)
- 20% - high dimensional content (our method)

Critical for estimating the effects for sparse signal—content similarity
Content & Context

- ~ 0.0 correlation between content and context novelty
- **30%** of hit probability captured by hypergraph of context + context
- Not linearly additive, but SUBSTANTIAL marginal effect
Scientists think through the complex network of content conditional on context
Science Think Differently from Scientists Think

Science thinks like a Global Bayesian
  ...by conditioning success/impact on affirmation of global priors

Scientists have much weaker priors

...but succeed by appearing to build on the shoulders of their audience

Negative crowd-sourcing - finding combinations unlikely to have been imagined nearly doubled the likelihood of success