ARBITER: Towards the Automated Detection of Technoscientific Emergence from Full-text Publications and Patents

Olga Babko-Malaya, Andy Seidel, Dan Hunter, Sean Stromsten, Fotis Barlos, Patrick Thomas, Marc Verhagen, Adam Meyers, James Pustejovsky, David C. Brock

September 25, 2013

BAE SYSTEMS

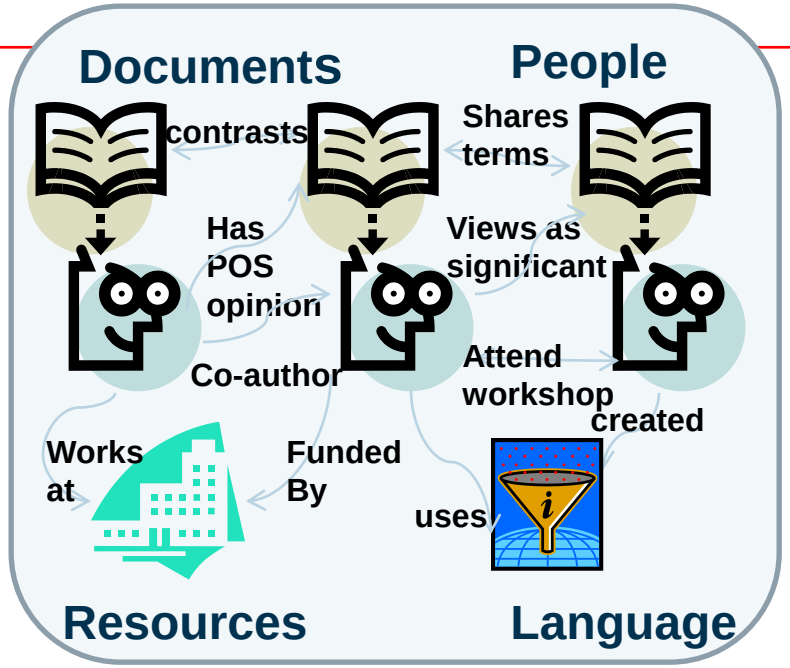
Acknowledgement

- Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.
- The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

BAE SYSTEMS

Applied Actant Network Theory

- Actant Network is a heterogeneous network of people and their physical and conceptual environment including their relationships and interconnections and *the linguistic artifacts of each.*
- In an Actant Network, individuals draw on a variety of conceptual, material, and sociocultural resources to transform existing domains within science and technology
- This intention-driven human activity causes networks to form or change



We model emergence by measuring the character and evolution of Actant Networks

- Robustness: diversity of funders, diversity of document types,…
- Growth: growth of positive opinion, growth of the size of the community,..
- Novelty: new researchers, new terms,..

BAE SYSTEMS

Definition of Emergence

- Emergence: *growth in the robustness of an actant network*.

- **Robustness:** the ability of the network to maintain its connectivity and continue to function, despite the failure or removal of crucial actants
  - More robust actant networks are more likely to persist and grow in the future
  - Robust networks exist not only in emerging domains, but also in mature technologies and scientific fields

- **Increasing Robustness:** differentiates emerging from mature domains
  - Networks associated with a mature domain are relatively stable

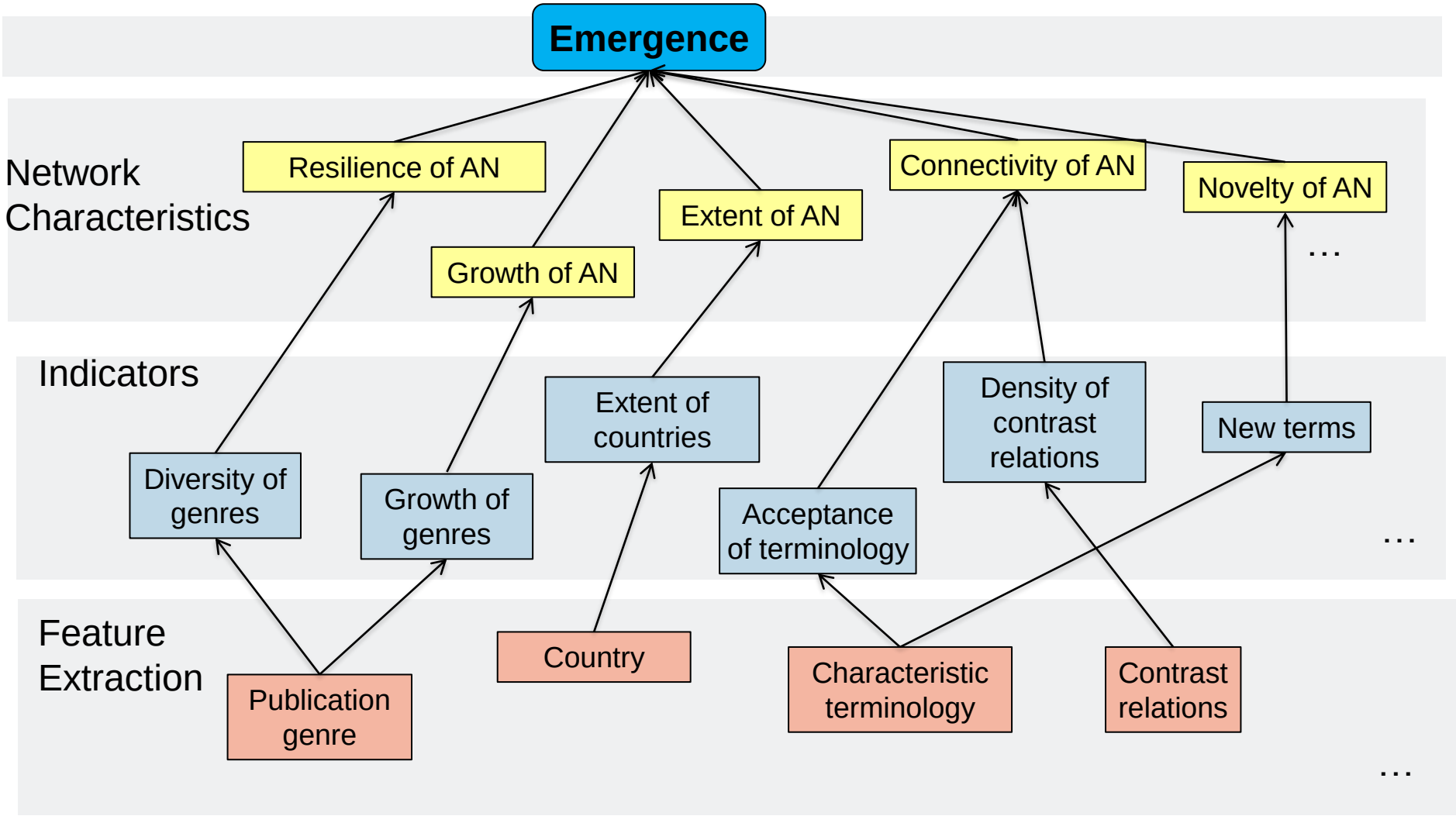
BAE SYSTEMS

Characteristics of Actant Networks

1. **Extent of Actant Networks**: measures the size of network elements
2. **Resilience**: measured by diversity of actant types
3. **Connectivity:** measures traffic and the amount of relationships
4. **Novelty**: measures burst of activities, lots of new stuff happening
5. **Growth**: measures growth of different elements of network
6. **Incommensurability**: groups of actors work in parallel isolation from each other
7. **Active Disagreement:** groups of actors manifest an active disagreement
8. **Topical Uncertainty**: multiple areas or items of topical uncertainties
9. **Extent of Industrial Involvement**: the prevalence of the marketplace actant
10. **Extent of Patenting:** there is a greater interest in patents
11. **Maturity of the Field:** measures the technology life cycle

BAE SYSTEMS

A Hierarchy of Indicators



BAE SYSTEMS

Metadata and Full Text Features

People:

- Researcher, Inventor, Patent Assignee
- Industrial Researcher

Resources

- Funder, Affiliation Organization, Organization, Country, Trademark
- Funder Type, Patent Assignee Type, Affiliation Organization Type

Publications

- Paper, Patent
- Patent Generality, Patent Originality, High Impact Paper, High Impact Patent

Language

- Topics, terminology, characteristic terminology
- Genre Classification: review paper, product review, debate, patent maturity score
- Document Structure
- Relations between terminology, citations, people, and organizations

BAE SYSTEMS

Relations between Terms and Documents

- ABBREVIATE
  - *highly pathogenic avian influenza virus **(**HPAIV**),** beta galactosidase **(**βgal**)***
- EXEMPLIFY
  - *cholesterol phenotypes **including** TC and HDL-C*
- ORIGINATE RELATIONS: DISCOVER, MANUFACTURE, SUPPLY
  - *Calcium chloride dihydrate was **supplied** by Merck*
- OPINION RELATIONS: POSITIVE, NEGATIVE, SIGNIFICANT, PRACTICAL, STANDARD
  - *We focused on the effect of IGFBP5 on HSC, **using** the human LX2 cell line* [PRACTICAL]
- RELATED_WORK RELATIONS: CO-CITATION, CONTRAST, CORROBORATION, BASED_ON, BETTER_THAN
  - *German Patent Application No. 100 53 373 **by contrast to** German Patent Application No. 197 35 624*

BAE SYSTEMS

Document Genre Classification

- Genre is a measure for the kinds of texts over diverse topics or fields
  - Abstract, Article (Proceedings Paper, Research Article, Review Article), Biographical, Case Report, Commentary, Correction, Discussion, Editorial, Letter, Meeting Report, News, Patent Report, Report, Review (Book Review, Other Review, Product Review), Short Communication
  - f-score on genres is 0.72

- Debate score: encodes whether the function of the document is to engage in a debate
  - f-score on the English debate feature is 0.78

- Scientific language score: encodes how similar a document/patent is to scientific articles

BAE SYSTEMS

Phase 1 Evaluation: FUSE Challenge Questions

- Challenge Questions:
  - Was there a community of practice around <concept> during <time period>?
  - Was a practical (vs. theoretical) application of <concept> demonstrated during <time period>?
  - Were there debates within the scientific community on <concept> during <time period>?

- Data for evaluation included SME's judgments to these Challenge questions for 8 technologies and 6 time periods:
  - Technologies:
    - DNA Microarrays; Genetic Algorithms; Cold Fusion; Steganography; RF Metamaterials; Horizontal Gene Transfer; Tissue Engineering; and RNA Interference
  - Time periods:
    - 1981-1985; 1986-1990; 1991-1995; 1996-2000; 2001-2005; 2006-2010.

BAE SYSTEMS

Promising Indicator Hypotheses

Community of Practice	Practical Application	Debate
1. Diversity of Genres	1. Growth of Practical relations	1. Debate-type articles score
2. Extent of Countries	2. Growth of Patents	2. Patent silos augmented with contrast relations
3. Extent of Organizations	3. High Impact patents	3. Extent of contrast relations in Abstract section
4. Extent of New Researchers	4. Percentage of Industrial researchers	4. Extent of contrast relations in Methods and Results section
5. Extent of Researchers	5. Density of Practical relations	5. Density of contrast relations
6. Density of Abbreviations	6. Density of Originate relations	6. Scientific articles silos augmented with contrast relations
7. Density of Exemplify	7. Growth in the Percentage of Commercial Funders	…
8. Diversity of Funders	8. Growth in the Percentage of Commercial orgs	
9. Density of Opinion: Positive	9. Average Patent maturity score	
10. Extent of High Impact patents	…	
11. Percentage of Review Type articles		
12. Extent of Funders		
...		

- Indicators are ranked based on the estimated mutual information of each indicator with the ground truth data

BAE SYSTEMS

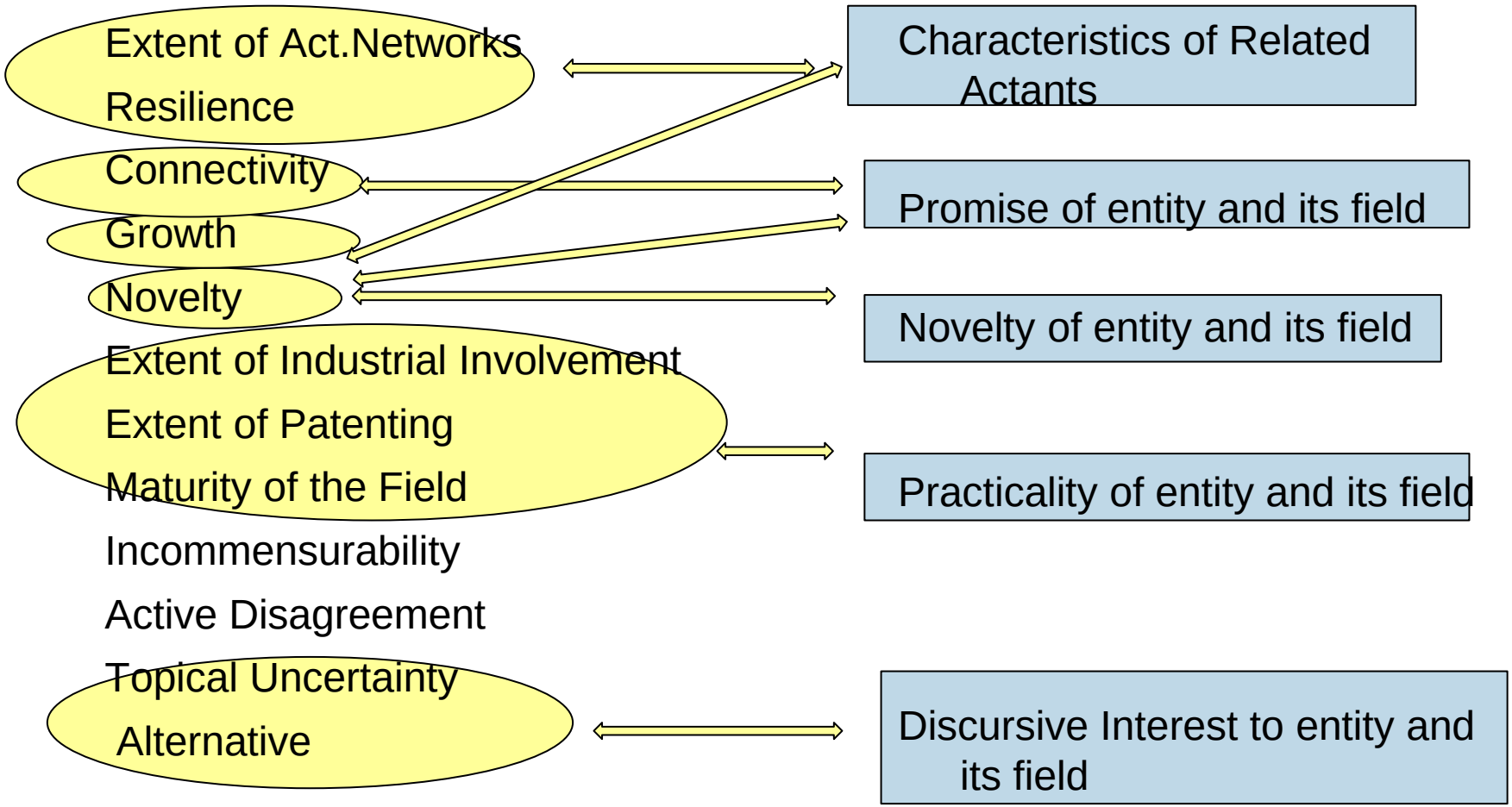
Example of Network Characterization: Tissue Engineering

	Extent	Resilience	Connectivity	Novelty	Growth	Industry	Patents	Maturity
1981-1985	Low	Mod	Mod	FALSE	0-0.25	FALSE	FALSE	FALSE
1986-1990	Low	Mod	High	TRUE	0-0.25	TRUE	TRUE	FALSE
1991-1995	Mod	Mod	High	TRUE	0-0.25	TRUE	TRUE	FALSE
1996-2000	High	Mod	High	TRUE	0-0.25	TRUE	TRUE	TRUE
2001-2005	High	Mod	High	TRUE	0.25-0.5	TRUE	TRUE	TRUE
2006-2010	High	High	High	FALSE	0-0.25	TRUE	TRUE	TRUE

- A small new community in the initial time period
- Fastest growth in 2001-2005, slowing down in 2006
- Significant Industrial Involvement and Patenting since 1986
- A mature and stable field in later time periods with High Extent, Resilience, Connectivity, and low Novelty

Phase 2: Prediction of future prominence of documents, terms, authors, and organizations

BAE SYSTEMS



BAE SYSTEMS

Conclusion

- Guided by Actant Network Theory, we developed a set of features, indicators, and network characteristics relevant to emergence

- A large variety of features and indicators offers insights into the character and evolution of emerging fields on different dimensions of emergence