Linguistic Corpus and Ontology for Comparative Analysis of Networks in International Development

Armando Razo [1] Markus Dickinson [2]

[1]Indiana University, Department of Political Science

[2]Indiana University, Department of Linguistics

September 29, 2014

Outline

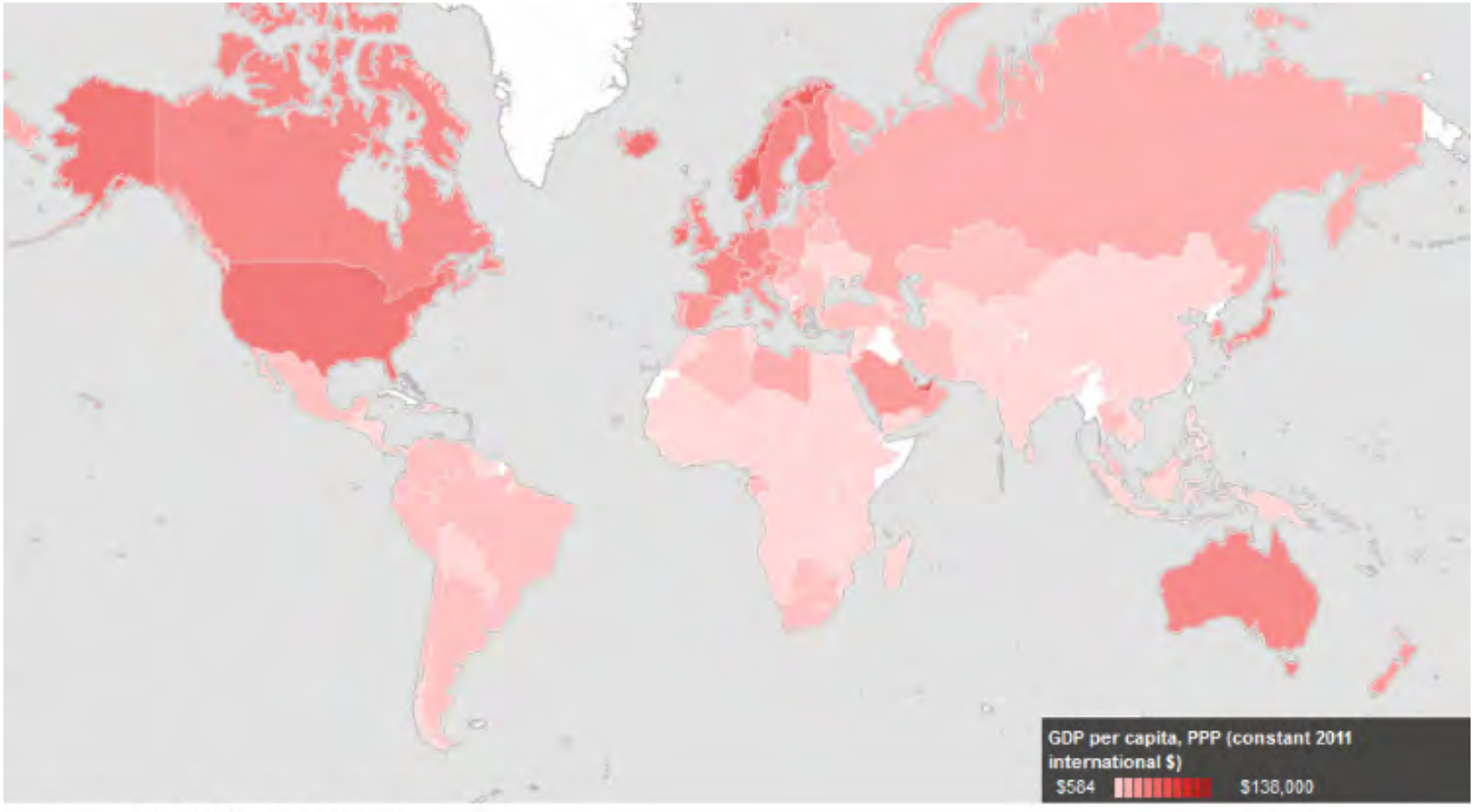
Scientific Domain and Network Science Challenges

Overview of Pilot Corpus and Web Ontology

Computational Linguistics Solutions

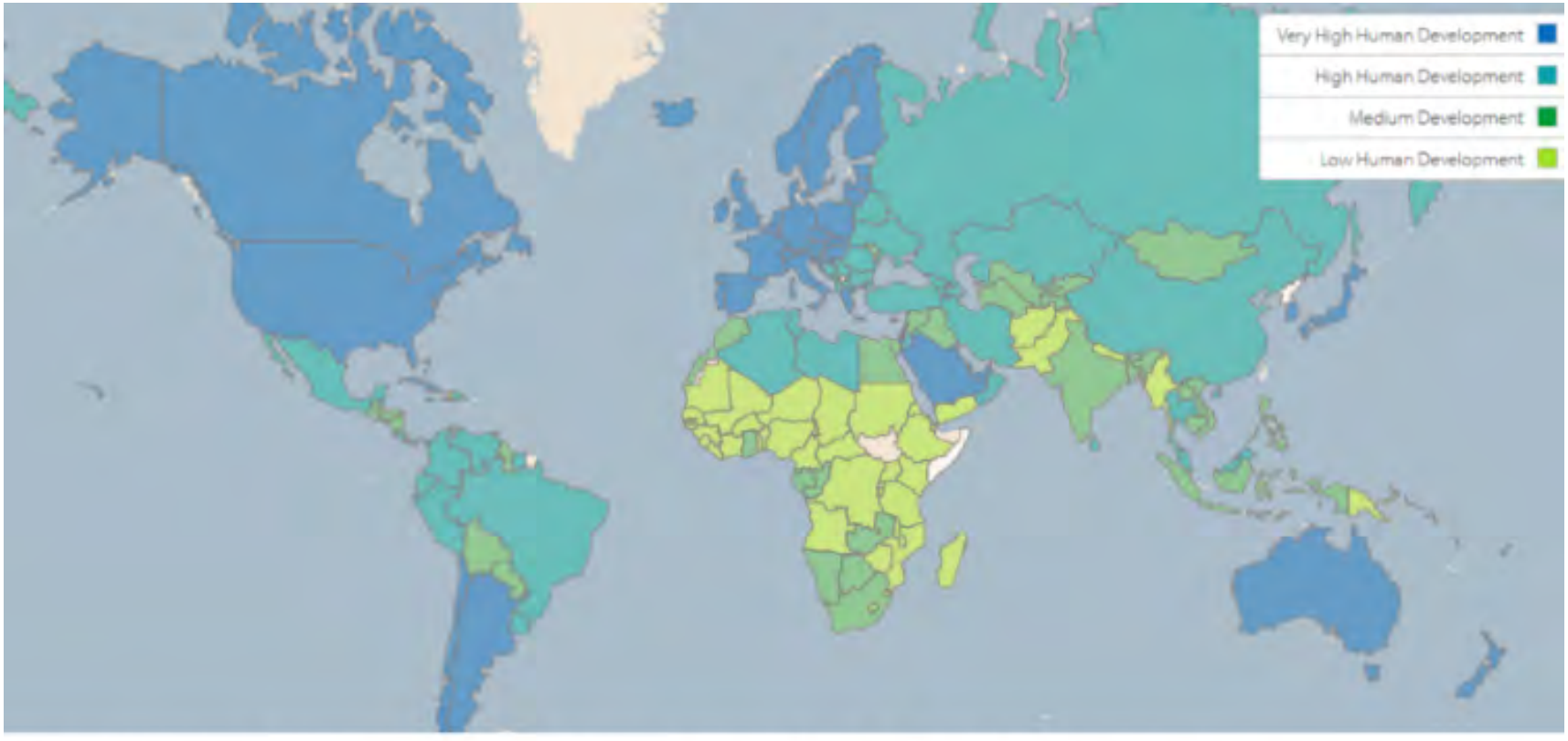
BIG Question of Social Science

Why are some countries rich and (most) others poor?



Source: http://data.worldbank.org.

Variable economic performance affects human development



Source: United Nations Development Programme (2014). Human Development Report. http://http://hdr.undp.org/en/countries.

Whether a country is rich or poor greatly impacts poverty and inequality, public health, and many other developmental outcomes.

Our answers to the BIG question are limited

- Scholars and policymakers claim that **"institutions" matter**.
  - Rule-based governance
  - Formal institutions of limited government
- But the conventional wisdom is oftentimes ineffective
  - Institutions don't always work or work differently
  - Sometimes the lack of prescribed institutions produces good results (China)

The developing world appears to be relational

- The study of international development has established the high impact of pervasive **informal** institutions.
- Not sure what these are exactly, but entail various types of relations such as:
  - social interactions
  - social relations
  - political connections
  - non-programmatic policies

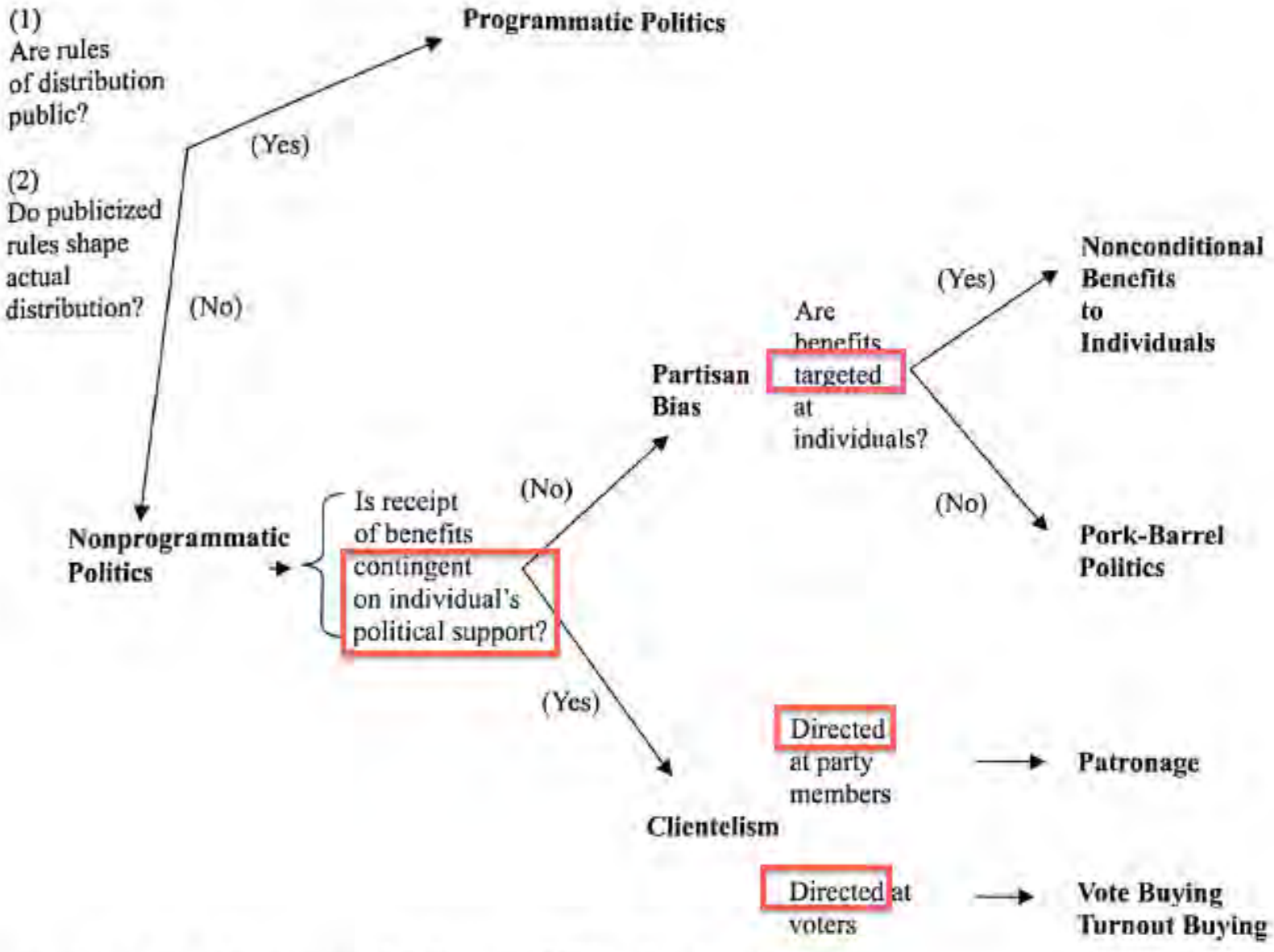
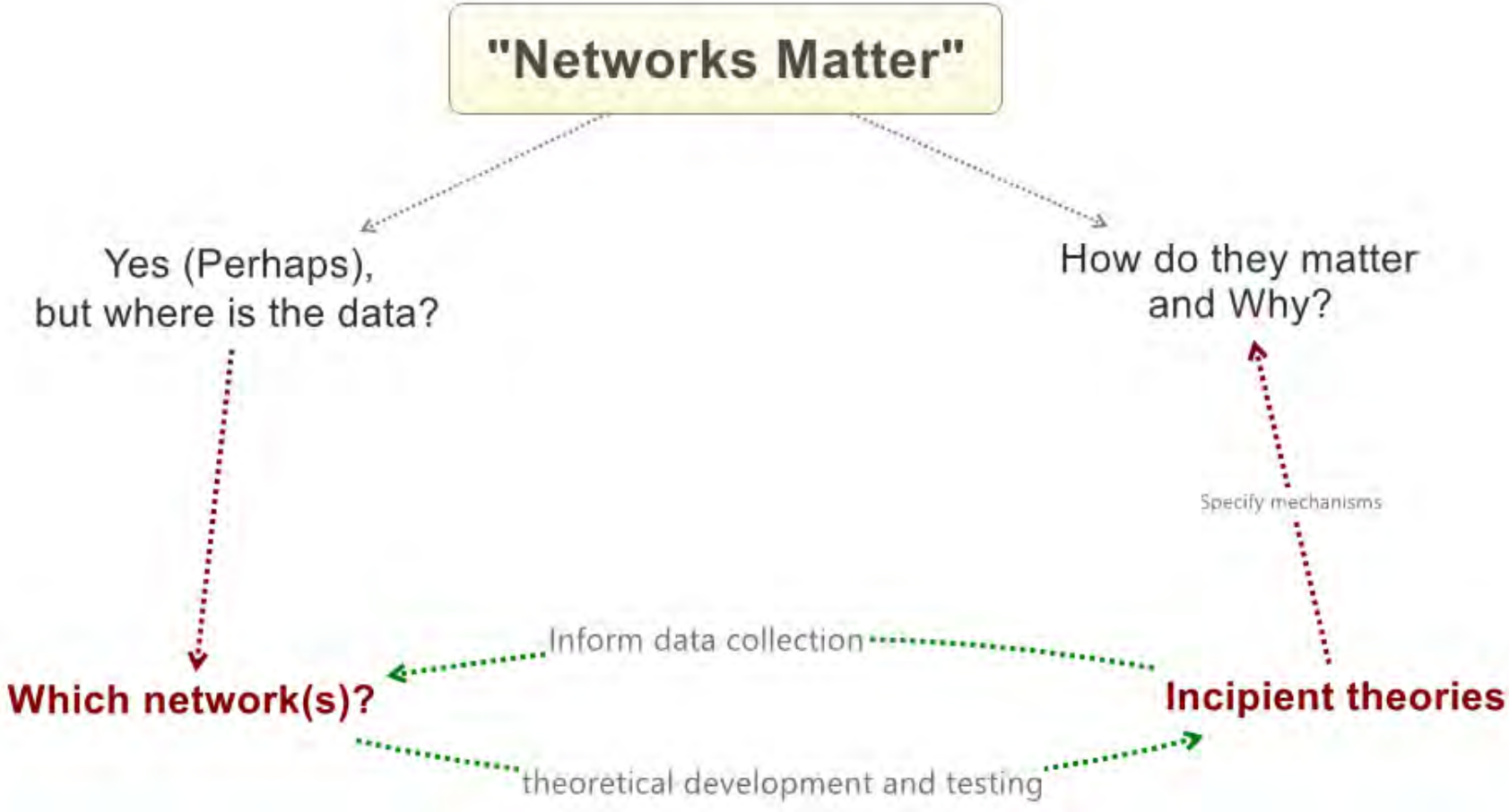


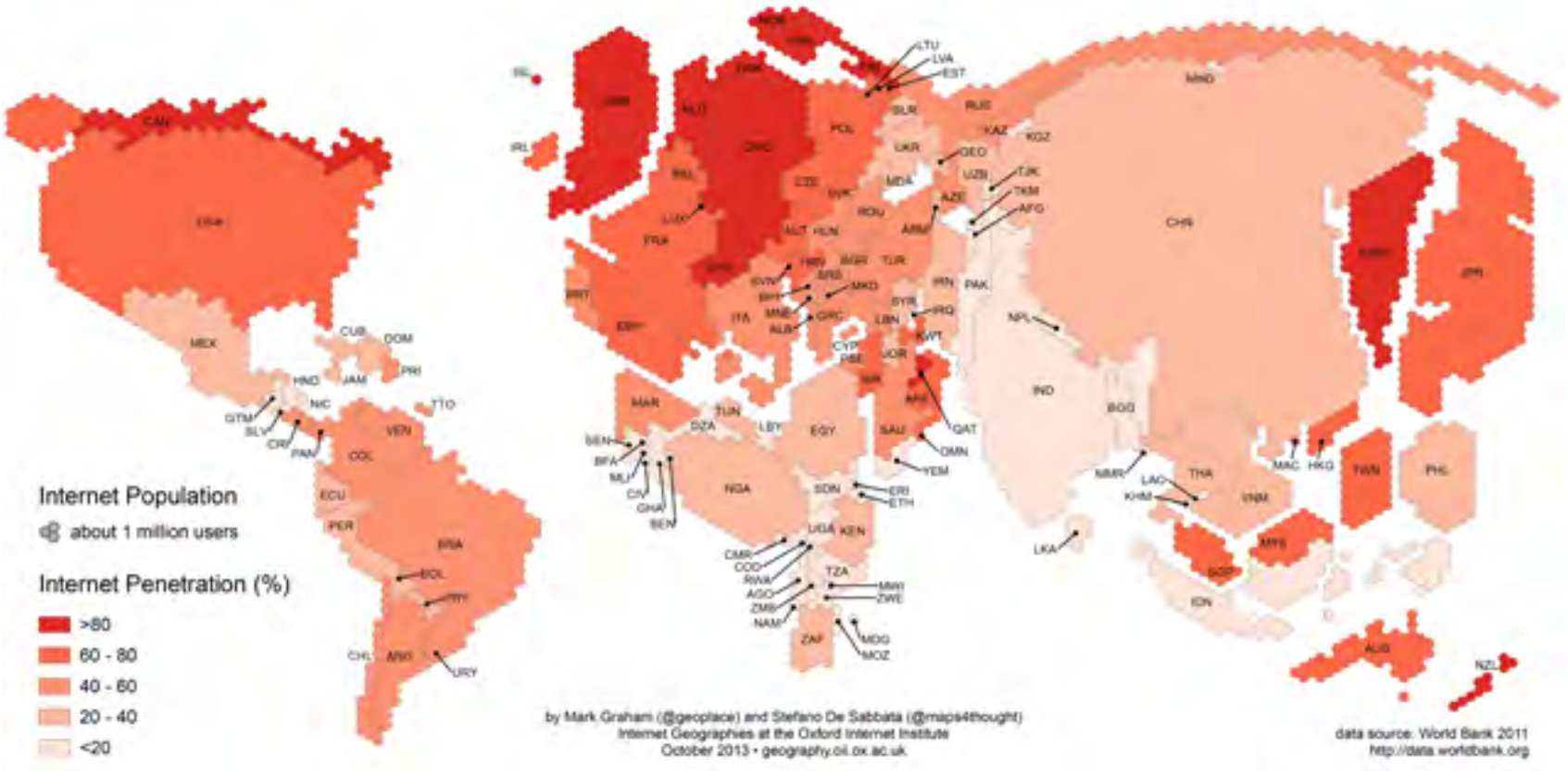
FIGURE 1.1. A Conceptual Scheme of Distributive Politics.
Source: Stokes et al (2013). Brokers, Voters, and Clientelism. Cambridge University Press.

It's now common to prescribe relational solutions

- Constructive: "Social capital," public-private partnerships, etc.
- Destructive: mitigate clientelism, corruption, etc.


"Networks Matter"
Yes (Perhaps),
but where is the data?
How do they matter
and Why?
Specify mechanisms
Inform data collection
Which network(s)?
Incipient theories
theoretical development and testing

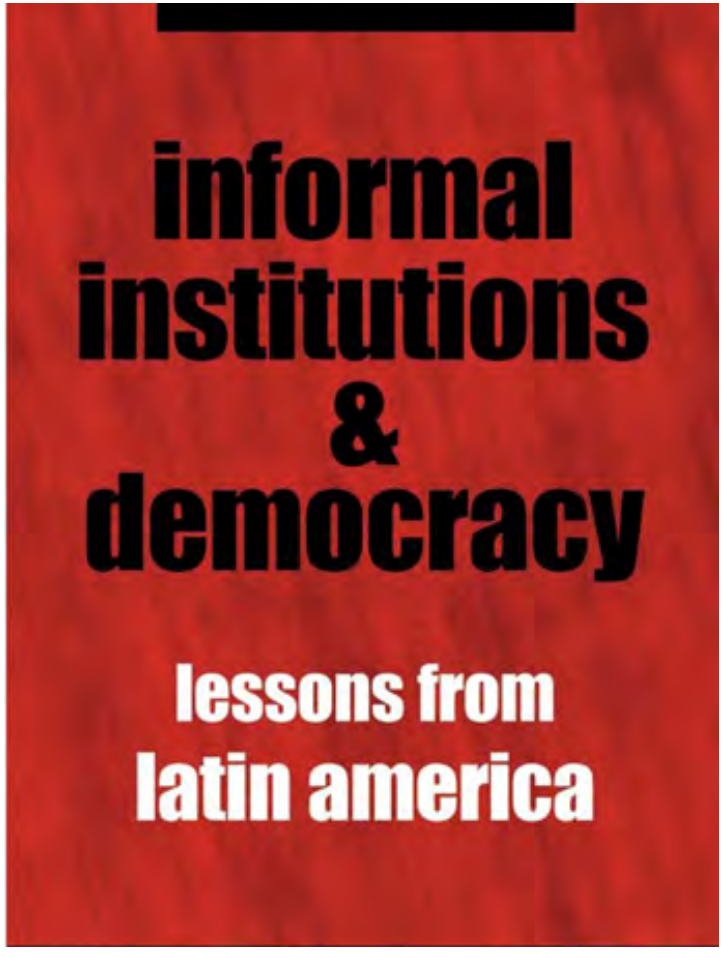
Uneven Internet Population and Penetration → Less Data



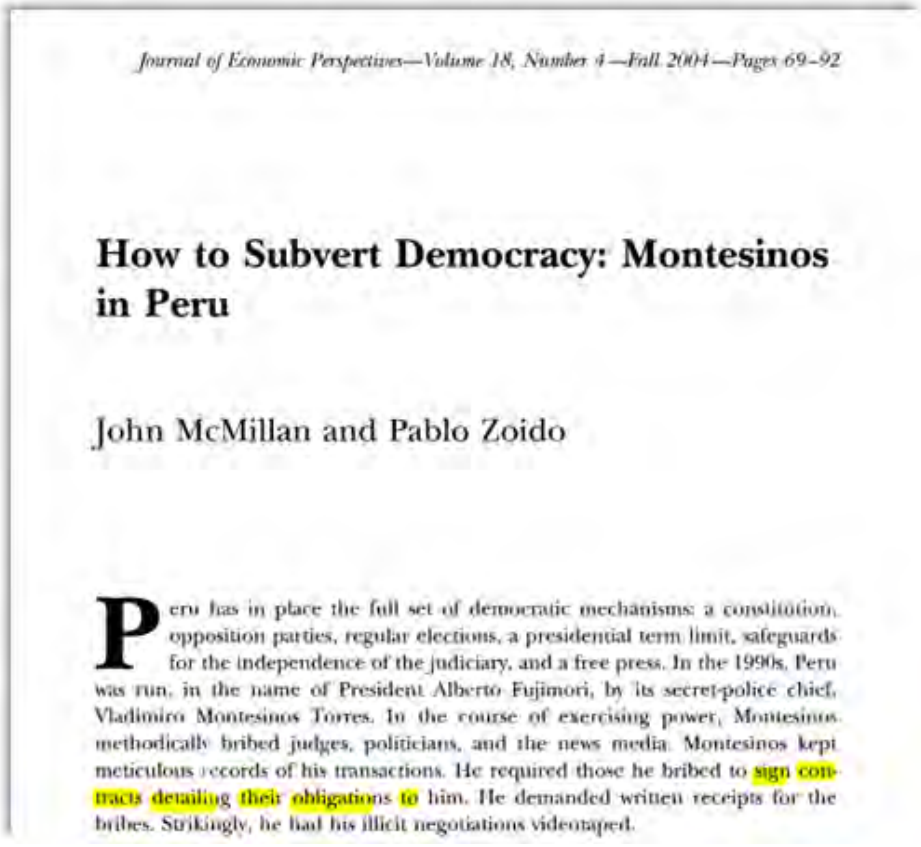
How to jump start a network science of international development

- Multiple challenges:
  - Highly interdisciplinary $\rightarrow$ Lack of common descriptive framework and methodological standards
  - Theoretical confusion about the role of networks
  - Very limited relational data
- Two-pronged **technological** solution with two guiding principles:
  - Don't reinvent the wheel: extract existing knowledge
  - Make it easy to do network analysis proper: provide an ontological framework to identify networks

Corpus: sample books



Corpus: sample articles

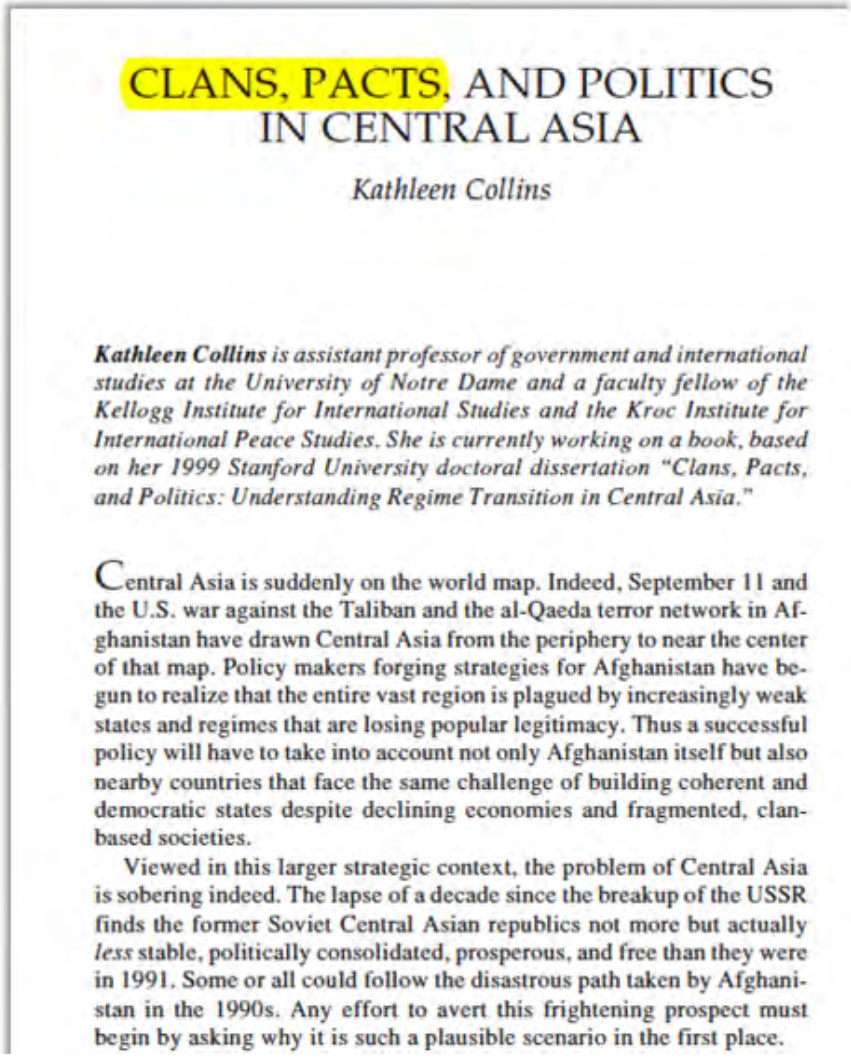


Journal of Economic Perspectives—Volume 18, Number 4—Fall 2004—Pages 69–92

How to Subvert Democracy: Montesinos in Peru

John McMillan and Pablo Zoido

Peru has in place the full set of democratic mechanisms: a constitution, opposition parties, regular elections, a presidential term limit, safeguards for the independence of the judiciary, and a free press. In the 1990s, Peru was run, in the name of President Alberto Fujimori, by its secret-police chief, Vladimiro Montesinos Torres. In the course of exercising power, Montesinos methodically bribed judges, politicians, and the news media. Montesinos kept meticulous records of his transactions. He required those he bribed to sign contracts detailing their obligations to him. He demanded written receipts for the bribes. Strikingly, he had his illicit negotiations videotaped.



CLANS, PACTS, AND POLITICS IN CENTRAL ASIA

Kathleen Collins

***Kathleen Collins** is assistant professor of government and international studies at the University of Notre Dame and a faculty fellow of the Kellogg Institute for International Studies and the Kroc Institute for International Peace Studies. She is currently working on a book, based on her 1999 Stanford University doctoral dissertation "Clans, Pacts, and Politics: Understanding Regime Transition in Central Asia."*

Central Asia is suddenly on the world map. Indeed, September 11 and the U.S. war against the Taliban and the al-Qaeda terror network in Afghanistan have drawn Central Asia from the periphery to near the center of that map. Policy makers forging strategies for Afghanistan have begun to realize that the entire vast region is plagued by increasingly weak states and regimes that are losing popular legitimacy. Thus a successful policy will have to take into account not only Afghanistan itself but also nearby countries that face the same challenge of building coherent and democratic states despite declining economies and fragmented, clan-based societies.

Viewed in this larger strategic context, the problem of Central Asia is sobering indeed. The lapse of a decade since the breakup of the USSR finds the former Soviet Central Asian republics not more but actually *less* stable, politically consolidated, prosperous, and free than they were in 1991. Some or all could follow the disastrous path taken by Afghanistan in the 1990s. Any effort to avert this frightening prospect must begin by asking why it is such a plausible scenario in the first place.

(a) One-to-one

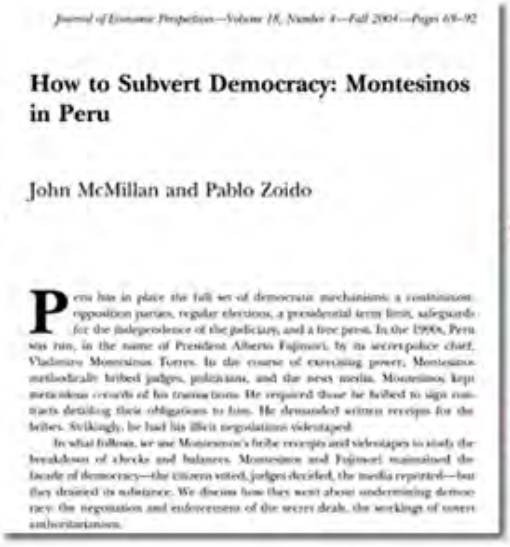
(b) One-to-many

CLANS, PACTS, AND POLITICS IN CENTRAL ASIA

Kathleen Collins

***Kathleen Collins** is assistant professor of government and international studies at the University of Notre Dame and a faculty fellow of the Kellogg Institute for International Studies and the Kroc Institute for International Peace Studies. She is currently working on a book, based on her 1999 Stanford University doctoral dissertation "Clans, Pacts, and Politics: Understanding Regime Transition in Central Asia."*

Central Asia is suddenly on the world map. Indeed, September 11 and the U.S. war against the Taliban and the al-Qaeda terror network in Afghanistan have drawn Central Asia from the periphery to near the center of that map. Policy makers forging strategies for Afghanistan have begun to realize that the entire vast region is plagued by increasingly weak states and regimes that are losing popular legitimacy. Thus a successful policy will have to take into account not only Afghanistan itself but also nearby countries that face the same challenge of building coherent and democratic states despite declining economies and fragmented, clan-based societies.



Journal of Economic Perspectives—Volume 18, Number 4—Fall 2004—Pages 69–92

How to Subvert Democracy: Montesinos in Peru

John McMillan and Pablo Zoido

Peru has in place the full set of democratic mechanisms: a constitution, opposition parties, regular elections, a presidential term limit, safeguards for the independence of the judiciary, and a free press. In the 1990s, Peru was run, in the name of President Alberto Fujimori, by its secret-police chief, Vladimiro Montesinos Torres. In the course of exercising power, Montesinos methodically bribed judges, politicians, and the news media. Montesinos kept meticulous records of his transactions. He required those he bribed to sign contracts detailing their obligations to him. He demanded written receipts for the bribes. Strikingly, he had his illicit negotiations videotaped.

In what follows, we use Montesinos's bribe receipts and videotapes to study the breakdown of checks and balances. Montesinos and Fujimori maintained the facade of democracy—the citizens voted, judges decided, the media reported—but they drained its substance. We discuss how they went about undermining democracy: the negotiation and enforcement of the secret deals, the workings of totalitarianism.

Network type 1

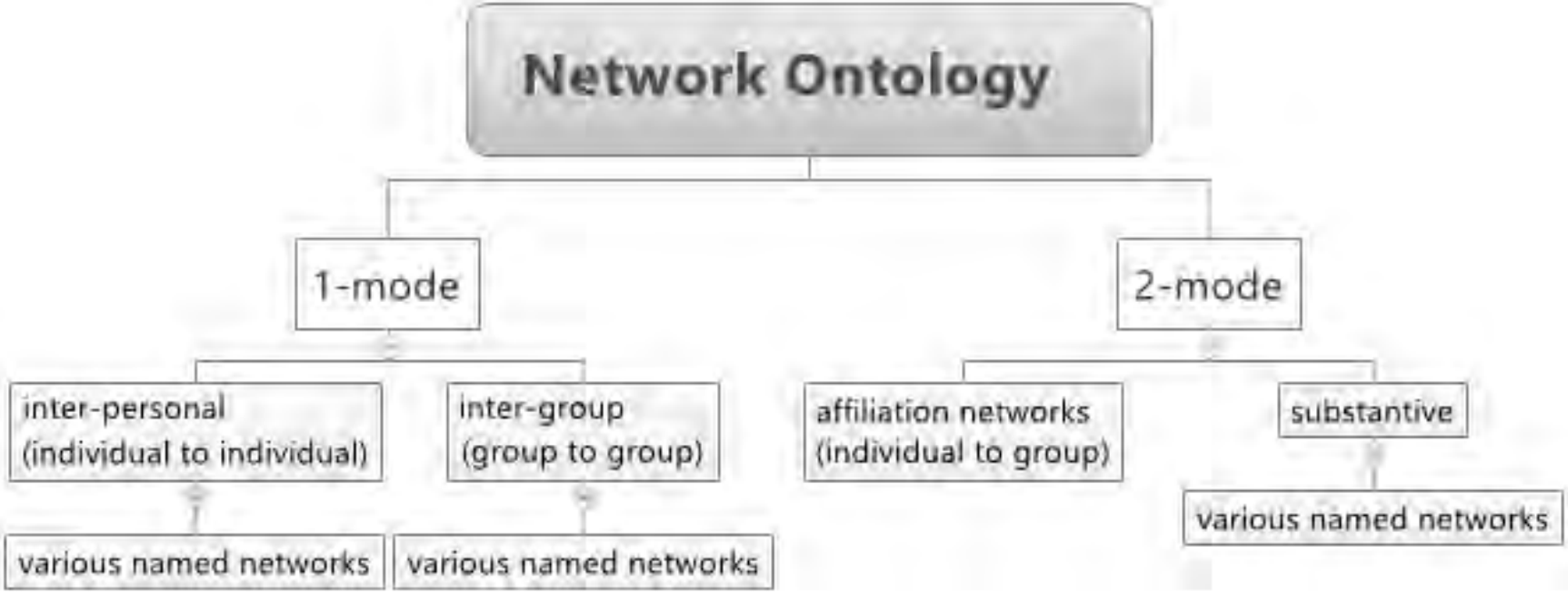
Network type 2

Network type 3

Network type *T*

What is an ontology?

- "Ontologies are content theories about the sorts of objects, properties of objects, and relations between objects in a specified domain of knowledge." (Chandrasekaran and Benjamins 1999).
- Basically, an ontology is a "list of classes" to define objects in a given domain (Noy and McGuinness 2005).
- We are developing three related (sub-)ontologies:
  - networks
  - developmental outcomes
  - research studies (corpus)


Network Ontology
1-mode
2-mode
inter-personal
(individual to individual)
inter-group
(group to group)
affiliation networks
(individual to group)
substantive
various named networks
various named networks
various named networks

Computational Linguistics Solutions

Q: How can we go from text ...

> *Policy actors seek network contacts to improve individual payoffs in the institutional collective action dilemmas endemic to fragmented policy arenas. The risk hypothesis argues that actors seek bridging relationships (well-connected, popular partners that maximize their access to information) when cooperation involves low risks, but seek bonding relationships (transitive, reciprocal relationships that maximize credibility) when risks of defection increase. We test this hypothesis in newly developing policy arenas expected to favor relationships that resolve low-risk dilemmas. ...*

to ontological terms?

- `self-organizing network`
- `game-theoretic partner selection`, ...

A: Natural Language Processing (NLP)!

Natural Language Processing

Natural Language Processing (NLP): "The goal of this . . . field is to get computers to perform useful tasks involving human language" (Jurafsky & Martin 2009, p. 1)

Applications include:

- conversational agents / dialogue systems
- machine translation
- question answering
- **information extraction**
- ...

What challenges does our task have for extracting structured information from unstructured data?

Challenges for NLP

Diversity of Data

The data covers a range of topics, written in different styles, from various academic fields, e.g.,

- Fowler & Jeon 2008: networks covering appellate courts,
- Collins 2001: Turkic, Persian, and Slavic ethnonational divisions

This leads to needing to spot important but low-frequency terms:

- "group members obtain ... information about ... the reputation, **indebtedness** and wealth of the applicant" (Atieno 2001)

Challenges for NLP

Ambiguity

Context-dependent definitions of networks & properties

- "...one question that arises is the extent to which **credit can be offered to the rural poor** to facilitate their taking advantage of the developing entrepreneurial activities." (Atieno 2001)
  - may indicate a network, but only if the supply of credit is contingent upon personal relations
- *association* may indicate PEOPLE-TO-PEOPLE, PEOPLE-TO-ORGANIZATION, or ORGANIZATION-TO-ORGANIZATION networks

Challenges for NLP

Shifting Reference

Each document may reference several networks, shifting between them

> *"There are a number of credit institutions that support small and microenterprise activities in the study region. . . . These include commercial banks, development finance institutions, NGOs, and rural credit organizations like SACCOs and ROSCAs. There are also a number of financial transactions taking place outside these institutions, like those between* **relatives and friends, traders, and welfare groups**." *(Atieno 2001)*

- shifts from ORGANIZATION-TO-ORGANIZATION to a PEOPLE-TO-PEOPLE network

Challenges for NLP

Extracting Network Features

Association between network mentions and its features may be spread far apart

- "**Each judicial citation** contained in an opinion is essentially a latent judgment about the case cited. ... We use the complete **network of citations** in all 30,288 majority opinions contained in the U.S. Reports from 1754 to 2002"
  - The network is of judicial citations

Solutions

Despite these challenges, NLP tools are good at giving us information from well-edited text

We currently have three components to our processing:

- Dependency parsing
  - To know, e.g., which features are truly connected to a network
- Relation & event extraction (not discussed today)
  - Re-use existing tools to determine who did what to whom
- Keyword filtering
  - Isolate linguistic structures that are relevant

Solutions

Keywords

Start by identifying a controlled set of vocabulary

- *network*, *system*, *actor*, etc.
- We will use these to filter out our linguistic information (next slide)

Next step: use a small set of initial seed terms and patterns to identify domain- or article-specific terms

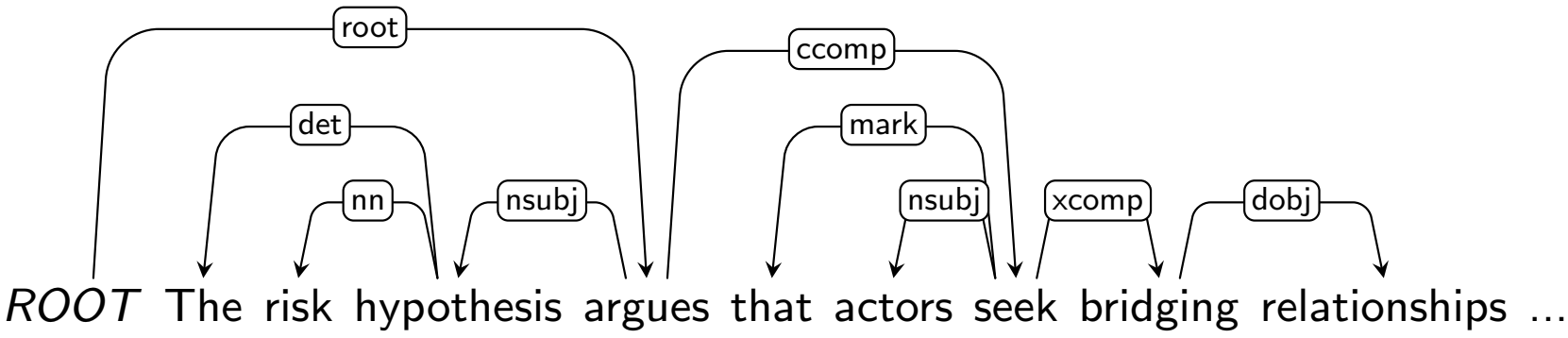
Solutions

Dependency Parsing

Parsers give some indication of who did what to whom

- Fairly fast & accurate for well-edited text

The start of a parse (from the Stanford Parser):



Even with some automatic error, note some things we can extract:

- *seek(actors,bridging)*
- relation between *bridging* and *relationships*

Solutions

Network features

By looking at what modifies a given keyword, we can extract various properties, e.g., for these keywords:

- network:
  - self-organizing
  - policy
- actor:
  - policy
  - popular
- relationship:
  - partners
  - bonding
  - transitive
  - reciprocal

Solutions

Basic relations

Ruling out relations which are never relevant:

- network: *none*
- actor:
  - seek(actors, bridging)
  - select(actors,partners)
  - seek(X, actors, [as] partners)
  - seek(actors, supportive [relationships])
  - trust(actors, partners)
- relationship:
  - maximize(relationship, credibility)
  - resolve(relationship, dilemma)
  - supportive(relationship, project)
  - seek(cooperation, relationship)

Next step: use document structure & other information within a section/paragraph to gain confidence in terms being relevant

- Also: move from sentence-level to document-level

Acknowledgements

Financial Support:

This research is made possible by a Faculty Research Support Program (FRSP) grant from the **Office of the Vice President for Research** at IUB.

Research Assistance:

- Wen Li (Linguistics)
- Luke Shimek (Political Science & SPEA)
- Dan Whyatt (Linguistics)