Developing Practical Solutions to Real World Problems by Going from Words to Networks

Jana Diesner, PhD
GSLIS & Dept of Computer Science
UIUC

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

Team

Jana Diesner
(Project Lead)
Assistant Professor
GSLIS
Computer Science

Amirhossein Aleyasin
Master Student
Computer Science

Chieh-Li “Julian” Chin
MS ’12 GSLIS
MCS ’13 Computer Science

Ming Jiang
PhD Student
GSLIS

Jinseok Kim
PhD Student
GSLIS

Shubhanshu Mishra
PhD Student
GSLIS

‘Shadi” Rezvaneh Rezapour
Master Student
GSLIS

Kiumars Soltani
PhD Student
Informatics

Liang Tao
Master Student
Agricultural Engineering

Impact Assessment: A Story of What Foundations Want, Practitioners Do and Academics Study

- Philanthropic foundations give billions in funding to:
  - “Work with visionaries on the frontlines of social change worldwide” (Ford Foundation)
  - Create “informed and engaged communities” (Knight)
  - “Tackle critical problems” in a way that “emphasizes collaboration, innovation, risk-taking, and, most importantly, **results**” (Gates)
- Results: **Impact, i.e. change** w.r.t. knowledge, behavior, beliefs, … (Barrett & Leddy 2008, Napoli, 2014, Chattoo & Das, 2014)

Impact Assessment: A Story of What Foundations Want, Practitioners Do and Academics Study

- **Status Quos**:
  - Quantitative: frequency counts (the more the better) plus
  - Qualitative: small-scale, in-depth focus groups interviews, quotes
  - Official mandates for systematic, foundation-wide approaches emerged over recent years
- **Filmmakers, Authors**: Storytelling (Rose 2012), useful to them:
  - Strategic allocation of limited resources for outreach and campaign work
  - Leverage existing social capital and discourse
- **Scientists:** psychological effects of media on individuals
- Similar trend in **Academia:**
  - Traditional evaluation of scholarly impact: impact = citation counts and metrics computed over counts (h-index, i-index) (Hirsch, 2005)
  - Recent: altmetrics (e.g. sharing of raw data, number of article views and downloads, references to scholarly work in traditional & social media) (Thelwall et al. Sugimoto, 2013)

Approach: Network Analysis and Text Mining

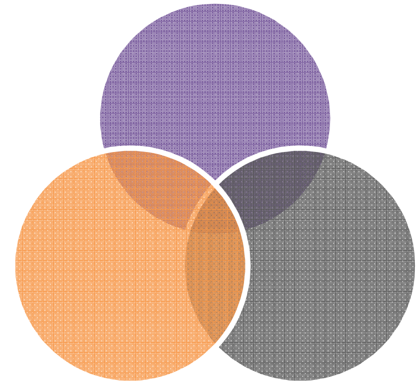
- **Question: How can we tell if an issue-focused media product has moved the needle on X?**
  - Empirical, rigorous, scalable, theoretically grounded
- **Assumption**: documentaries produced, screened, watched as part of **larger, dynamic ecosystems** of **stakeholders** and **information**
- **Operationalization**: identify, map, monitor, analyze **social networks** (of stakeholders) and **semantic information and networks** (of information) to study their structure, functioning and dynamics

DIMENSION	LEVEL		INDEX	ANALYTICS	ITEM
CONTENT	MESSAGE		Guiding Factor	Description Ranking weighing	Report by producers or funding agencies
	EXPECTED OUTCOME				
	EVALUATION PRIORITY				
	RESOURCE				
MEDIUM	RELEASE MEDIUM	OFFLINE	Outreach	Stats	Number of movies, CDs distributed Number of theatrical, Internet release Duration of release; Sales of product
		ONLINE			
	RESPONSIVE MEDIUM	MASS MEDIA	Mass Media Attention	Text Mining Web Analytics	Frequency of news coverage weighted by influence (article, opinion/editorial) Domestic, international broadcast
		USER MEDIA	User Media Attention	Text Mining Web Analytics Survey, Interview	Twitter, Facebook, Blogs, webpages Frequency of talking about, links included, user-created contents
		PROFESSIONAL MEDIA	Prestige		Number of festival acceptance Number of awards Number of professional reviews
		INTERPERSONAL INTERACTION	Intimate Attention		Conversation, talking on the phone or email, lectures, exchange of letters, etc.
TARGET	AUDIENCE SIZE		Reachability	Text Mining Web Analytics Archived Data Survey, Interview	Number of viewers or visitors
	HOMOGENEITY		Diversity		Geography & demography: location, age, gender, education, income
	AUDIENCE TYPE	SINKER	Passiveness	Text Mining Web Analytics Network Analysis	Number of inactive viewers
		TRANSMITTER	Leadership		Number of opinion leaders
	COLLECTIVE ENTITY		Advocacy	Text Mining Web Analytics Survey, Interview	Number of advocacy communities, colleges, schools, or NGOs
IMPACT	INDIVIDUAL COMMUNAL SOCIETAL GLOBAL	COGNITIVE	Awareness	Stats, Text Mining Web Analytics, Network Analysis	Frequency of names, ideas, thoughts, or concepts appeared in corpus Report of increased awareness
		ATTITUDINAL	Sentiment	Sentiment Analysis	Frequency of positive, negative, neutral sentiments of comments Personal, critics, mass media, and organizational responses Reaction to calls for action
		BEHAVIORAL	Engagement Enactment Connectedness Capacity Expansiveness Centralization	Text Mining Web Analytics Network Analysis	How well connected How much & far disseminated How centralized is the impact The route of diffusion Number of action pledges alliance and allied action of organization Discussion or decision by organizational, governmental, international policy/legislation makers sponsorship of bills, adoption, donation, funding, implementation, social movement or intervention
		TEMPORAL	Impact Dynamics	Longitudinal analysis	Comparison b/w multiple time points Duration of impact Increase vs. decrease Change vs. stability vs. reinforcement Introduction or shifts of topics Detection of social norm change

This is no computational fishing expedition. We have theory: CoMTI Framework

Diesner J, Kim J, Pak S (2014): Computational Impact Assessment of Social Justice Documentaries. *Journal of Electronic Publishing (JEP), special issue Metrics for Measuring Publishing Value: Alternative and Otherwise*

Workflow and Logic



Baseline
Public Discourse

- Social and semantic networks from meta data and content (relation extraction)
- Text summarization

Information Product

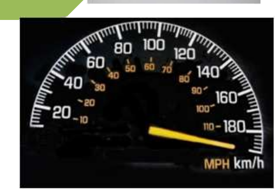
Public Discourse on Info Product

Change in Baseline

Theme

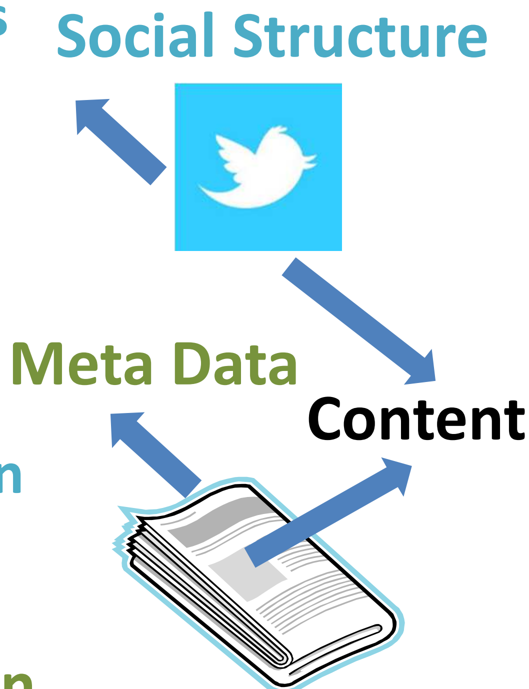
Technology: ConText

http://context.lis.illinois.edu



Technology: ConText

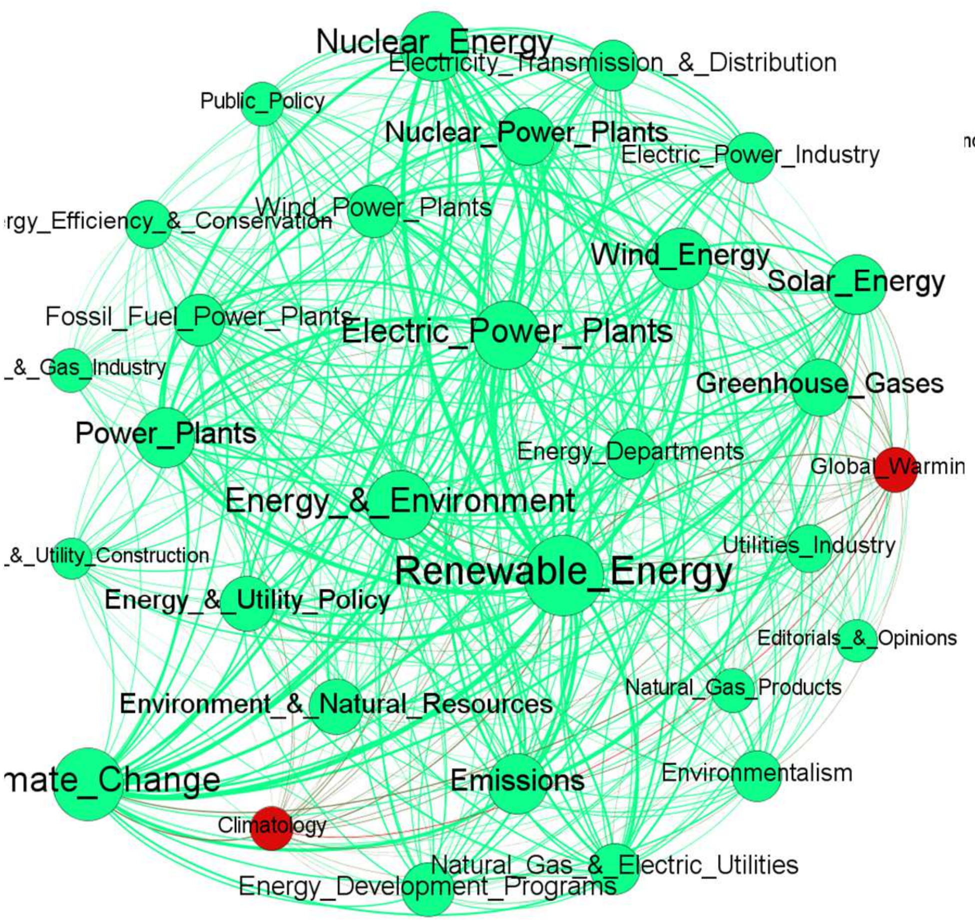
- Social Networks (FB, Twitter, YouTube)
- Semantic networks of content
- Tight integration with NodeXL
- Disambiguation
- Create meta-data databases
- Construct semantic networks

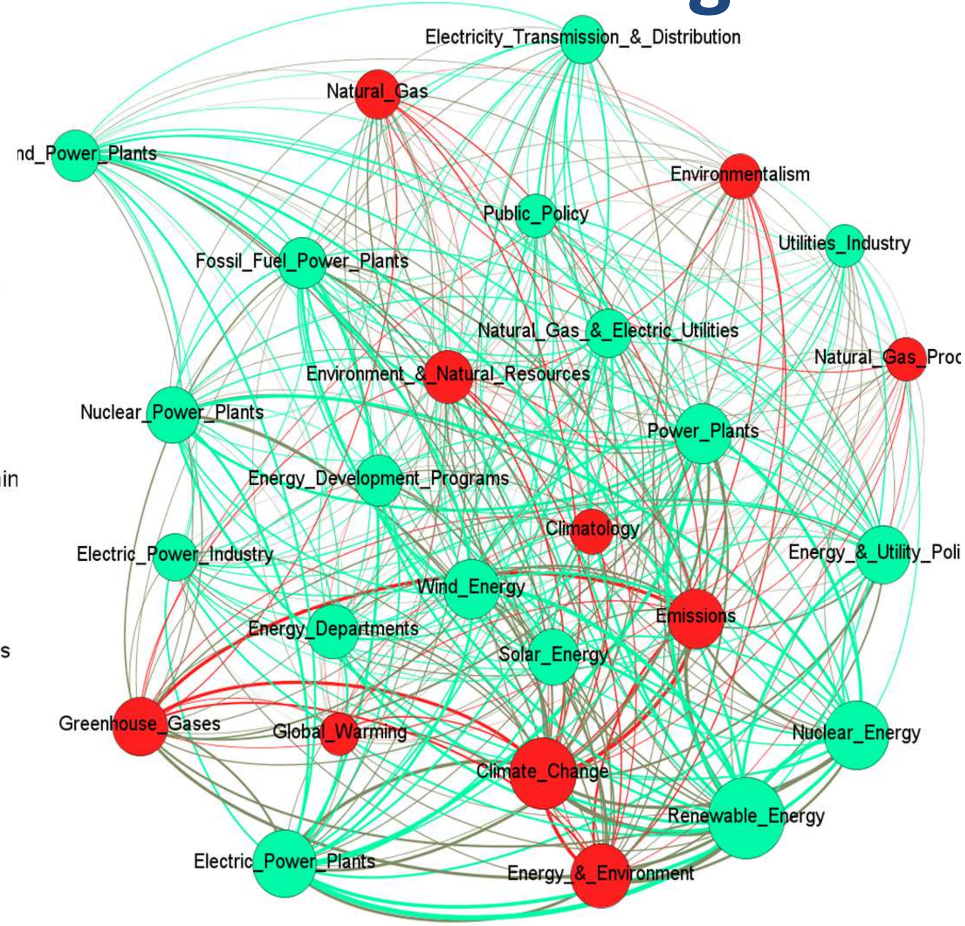


- **Text Mining & NLP:**
  – Pre-Processing
  – Summarization
  – Codebook construction and Application
  – **Entity Extraction**
  – **Relation Extraction**
    - **Co-occurrence**
    - **Semantics**
    - **Syntax**

	A	B	C	D	E	F	G	H
1	Source	Best Date	ıblication Ty	Title	Author	Geo	Organization	Person
2	The New York Times	2013-06-16	Newspaper	A Rebel Filmmaker Tilts (	By TOM ROSTON	EARTH (76%) UNITED STATES (79%)		
3	Wisconsin State Jou	2013-11-06	Newspaper	UW ALUM GENERATES	DOUG MOE , Wiscon	MADISON, WI, USA (73%) WISCONSIN, USA (91%); UTAH, U		
4	The New York Times	2013-06-12	Newspaper	Asking Environmentalists	By MANOHLA DARGI	TOHOKU, JAPAN (93%) JAPAN (93%)		PAUL ALLEN (50%)
5	Daily Variety	2013-01-23	Newspaper	Pandora's Promise	John Anderson	TOHOKU, JAPAN (79%); NEW YORK, USA (73%) JAPAN (79%		
6	Chicago Daily Herald	2013-06-14	Newspaper	Reel Life mini-review:		TOHOKU, JAPAN (79%) UNITED STATES (79%); JAPAN (79%		
7	The New York Post	2013-06-10	Newspaper	How they learned to stop	KYLE SMITH	NEW YORK, NY, USA (85%) TOHOKU, JAPAN (92%); NEW Y		
8	The Star Phoenix (S	2013-10-01	Newspaper	Film tackles the nuclear d	Scott Larson, The Star Phoenix			
9	The Oxford Times	2013-11-21	Newspaper	Parky at the Pictures (In	Parky at the Pictures	MIAMI, FL, USA (71%) CALIFORNIA, USA (76%); FLORIDA, U		
10	hollywoodreporter.cc	2013-08-12	Web Publicat	Paul Allen Lends Support	Gregg Kilday	UTAH, USA (92%) UNITED STATES (92		PAUL ALLEN (92%)
11	The Toronto Star	2013-07-12	NEWSPAPEI	What the world needs now is nukes				
12	Kamloops Daily New	2013-07-15	Newspaper	Pandora's white-hot deba	Bruce Cheadle, The C	TORONTO, ON, CANADA (72%) ONTARIO, CANADA (90%); T		

From Raw Data to Actionable Knowledge





Story of what Foundations Want, Practitioners Do and Academics Study

- **A dozen assessments later…**
- What?
  - Meta-review, lessons learned…
- So what?
  - Who cares beyond peer reviewers?
  - Usability for practitioners?
    - **Matching their standards?**
- Now what?
  - Practical implications

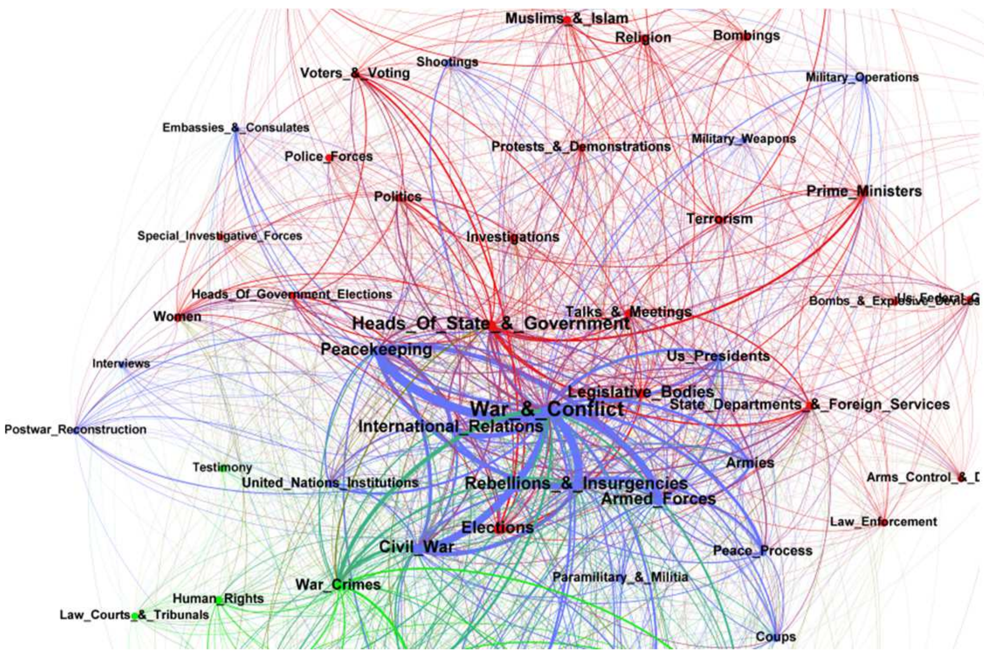
Case Study:
Can we capture what practitioners need?

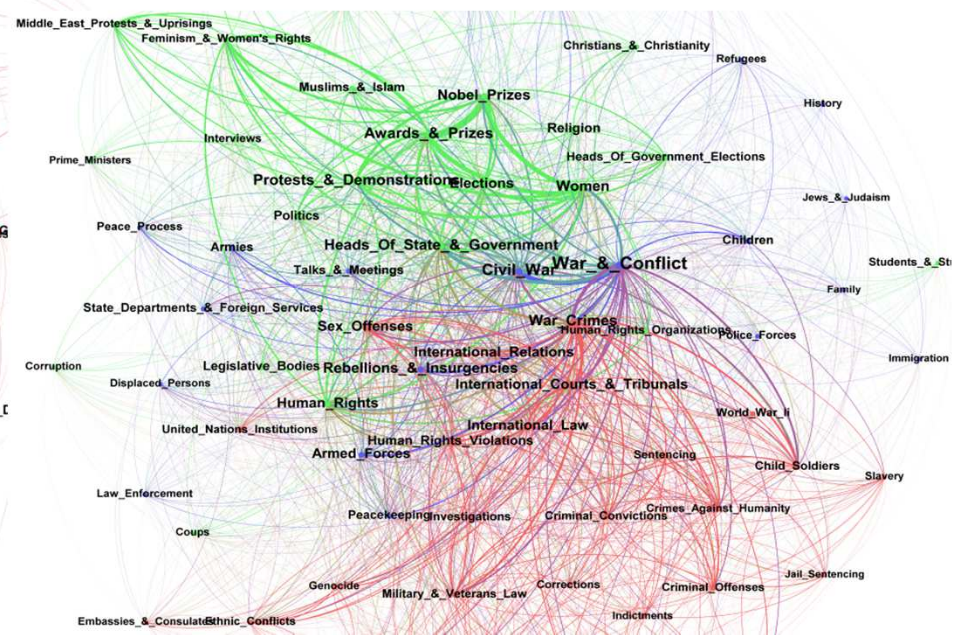
- "Women, War and Peace", impact report by Peace is Loud
  - Role of women in peace building in 4 geopolitical contexts
  - Quantitative: 12.57M viewers, 1,461 hostings of screenings
  - Qualitative: **census** of (social) media and screenings, interviews

Goal	Can we measure achievement?	How?
1. Build awareness for WWP	yes	Over-time semantic and social networks of media and social media data, additional natural language processing techniques (details in [6])
2. Spark dialogue	yes	
3. Reach and engage key constituencies	yes	
4. Continued utilization of series	yes	
5. Introduce series to new, varied audience	yes	
6. Increase public engagement with topic	partially (words yes, actions not)	
7. Inform stakeholders, serve as resource for stakeholders	not yet	
8. Highlight immediacy, proximity of topic	not yet	

Data, Networks (on theme)

Country	**Keywords** (baseline: woman, women, war, wartime, peace*, <country name>)	**Before**	**After**	**Press on film**
Afghanistan	peace talks, Taliban	450	1,069	4
Liberia	protest*, Charles Taylor	493	605	85
Colombia	gold*, displace* (not Olympic)	80	109	3
Serbia	rape, sexual violence	54	66	22



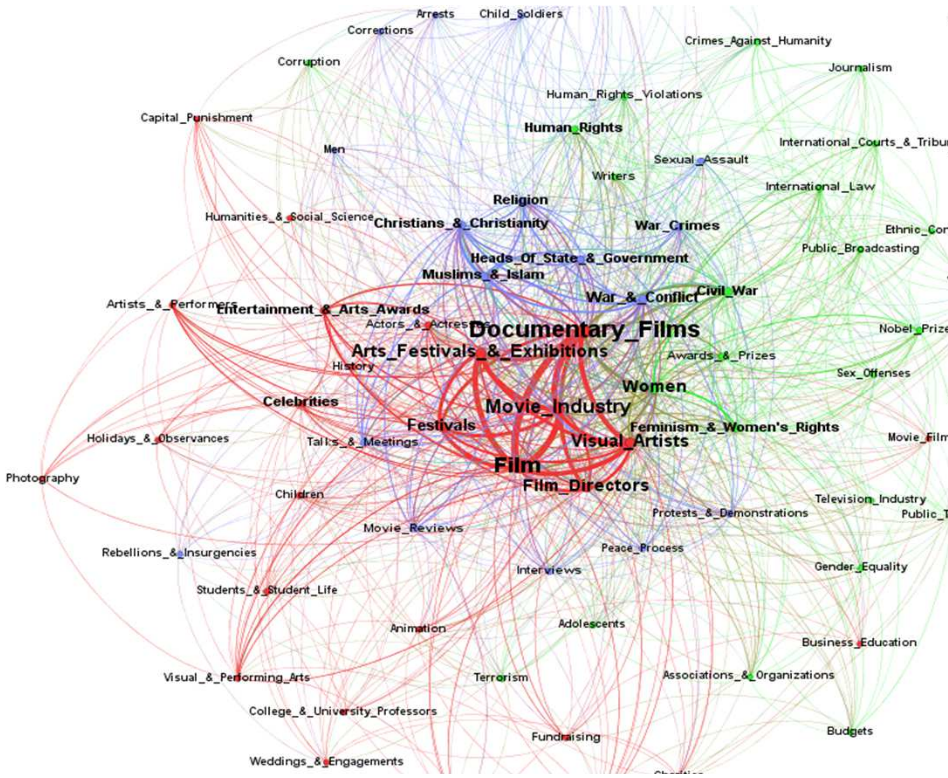


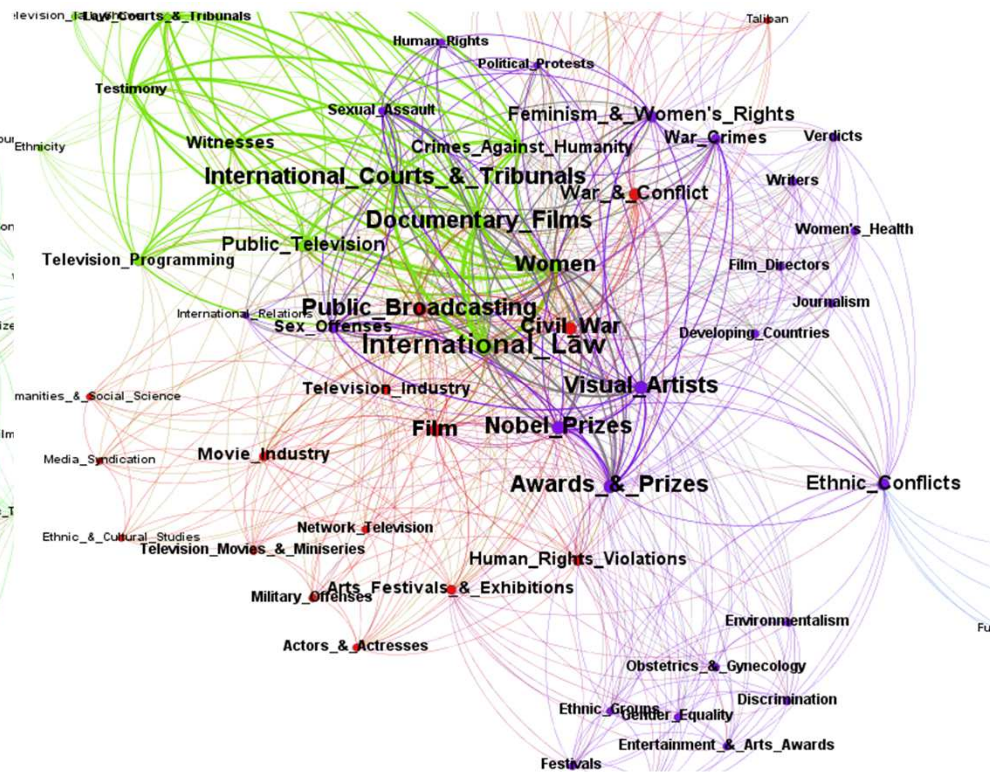
Salience of issue versus women

Film	Press on theme before release		Press on theme after release		Transcript
	Main cluster(s) and key nodes	**Women**	**Main cluster(s) and key nodes**	**Women**	(country name excluded)
Afghanistan (Peace unveiled)	(1) war & conflict, Taliban, muslims, peace process	2nd yet smaller cluster with human rights	(1) like before, (2) peace process, talks & meetings	marginal, separated from main clusters	women, Taliban, support, war, peace, conference
Liberia (Pray the devil back to hell")	(1) war & conf., civil war, rebellion & insurg. (2) elections	very marginal, no cluster	(1) like before (2) war crimes	3rd cluster with protests & demonstrations, nobel peace prize	Leyman Gbowee, women, peace, Charles Taylor
(Colombia (War we are living)	(1) war & conflict, human rights	marginal cluster with international relations	(1) rebellion & insurgencies, war & conflicts	2nd main cluster with human rights and displaced people	war, family, land, community, government
Serbia (I came to testify)	(1) war & conflict, ethnic conflict, religion (2) international legal issues	marginal cluster with sex offenses and human rights	(1) war & conflict, ethnic conflict, human rights (2) war crimes	marginal, no cluster	rape, women, witness, war, crime, tribunal

Networks (Press on Film)

- Liberia: film (making) and related festivals and awards, smaller cluster about religious issues
- Serbia: international legal matters related to women and violence
- All films: women more central in press on film than on topic





Conclusions and Expansion

With our impact assessment approach, one can:

- Measure achievement of large portion of common impact goals as defined by funders and evaluators
- Complement and enhance findings and interpretations obtained with standard techniques used by practitioners

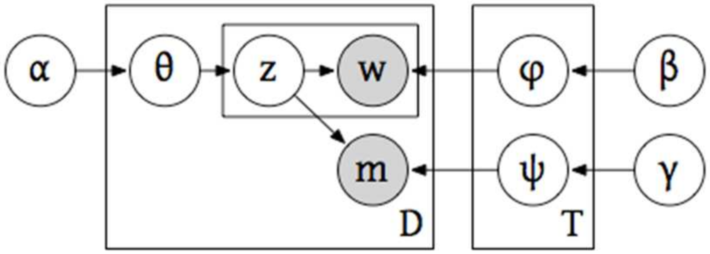
Current expansions (NCSA fellowship):

- Additional sources: Legal and corporate reports
  - Test for macro level impact: Political, Corporate, Cultural, Human Rights, Educational
- Prediction models for impact detection on 9 item scale:
  - Behavioral Change
  - Cognitive Change
  - Change in intensions
  - Emotional Change
  - Attitudinal Change
  - Active Reflection
  - Empathy
  - Summarization
  - Ranking or Rating

Case studies & Lessons Learned

- **Environment:**
  - **This Changes Everything**, by Naomi Klein, 2014
  - **Pandora's Promise**, by Robert Stone, 2013
- **Conflict/ Violence:**
  - **One Mile Away, by Penny Woolcock, 2013**
- **Race:**
  - **Through a Lens Darkly** (and the Digital Diaspora Family Reunion TV), by Thomas Allen Harris, 2013
- **Legal:**
  - **The House I live in**, by Eugene Jarecki, 2012
  - **A Kind of Order**, by Noel Schwerin, 2013
- **Education/ Human Rights:**
  - **Solar Mamas**, by Mona Eldaief & Jehane Noujaim, 2012
  - **Women War and Peace** (five-part television series), by PBS, 2011
- **Health:**
  - **Fed-Up**, by Stephanie Soechtig, 2014

Methodology:
Summarization (Topic Modeling)

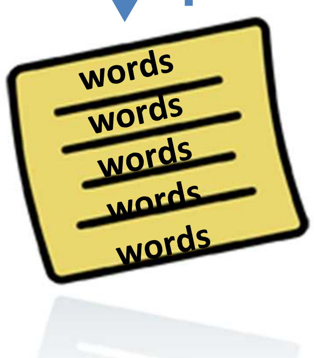


some latent structure, probabilistic graphical model

Theme 1 (70%)
Theme 2 (20%)
Theme 3 (10%)

Probabilistic Generative Process

Bayesian Inference



XXX **XXX** XXX XXX **XXX** XXX XXX
XXX **XXX** XXX XXX **XXX** XXX XXX
XXX XXX **XXX** XXX XXX **XXX** XXX
XXX XXX XXX XXX **XXX** **XXX** XXX
XXX **XXX** XXX XXX XXX XXX XXX

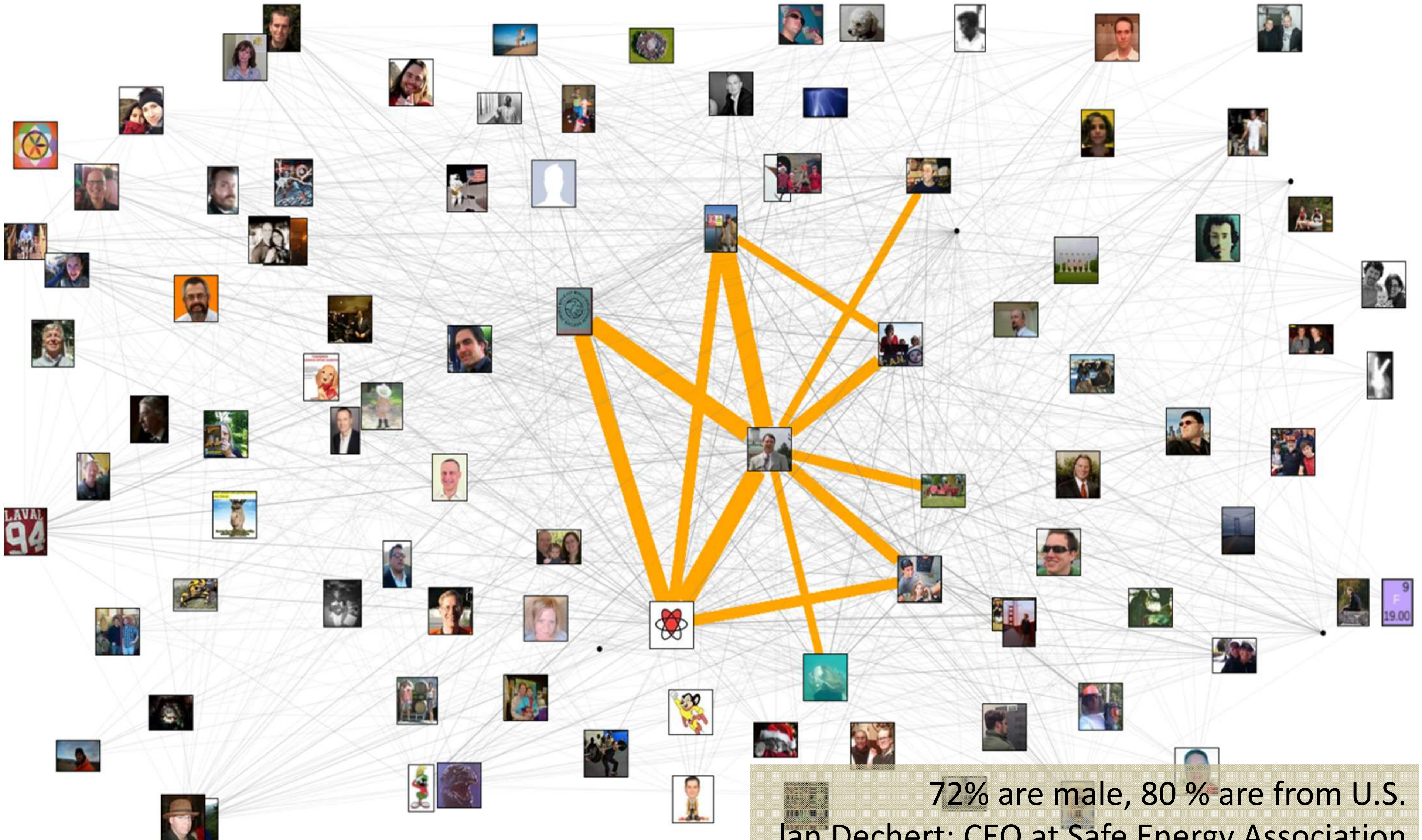
Comparing Substance/ Content of Baseline and Ground Truth (Solar Mamas)

BL: Press on theme: poverty in Arabic world & women, health, employment & development	
22%	health water people areas education government cent food poor
21%	development world years economic Arab poverty country time social
17%	women children work countries leaders time government world people
16%	women education empowerment girls women's gender war school child
13%	United Minister Education Development Nations Women
12%	President APRC people Oct election Development Jammeh support

GT: Transcript: storytelling (social conflict) and issue (training & employment for females)	
23%	back India don't **kids won't** I'm **call husband**
21%	work make **village** solar back women girls years
21%	**husband daughters** meeting things stay can't work girls
17%	**didn't role life** world trainees day India problem
17%	months mind **man mother** can't situation **sin** put

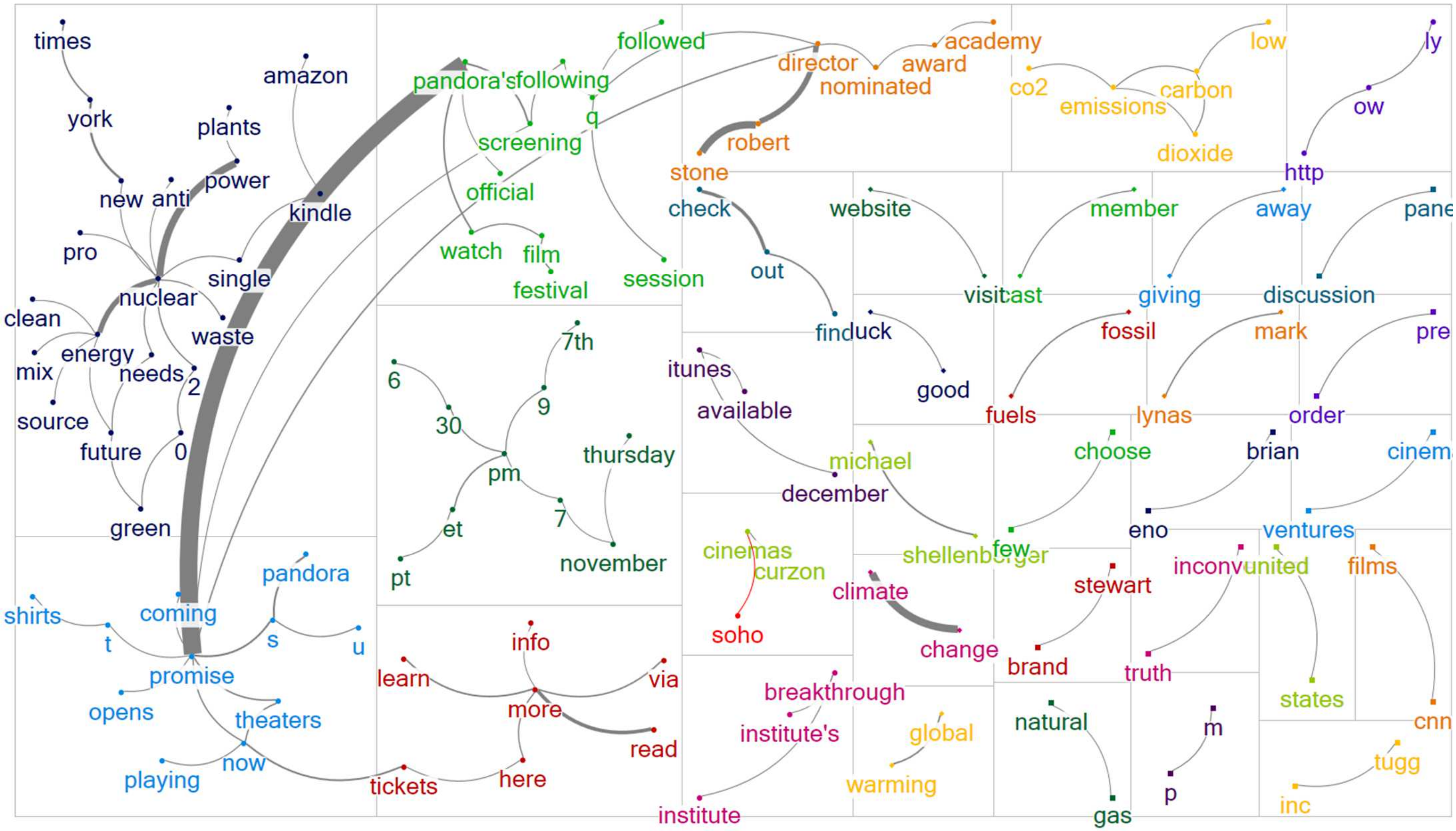
Press on documentary: poverty among people in the Arabic world, especially women	
93%	film **poverty** people **Arab** documentary films world **women**
3%	women solar Barefoot India College home back train
2%	p.m Free Ave Film National Center a.m Park
2%	Rafea Solar it's story Mamas mother Jordanian husband

"Public Opinion" on Social Media: Social Network of Co-commentors

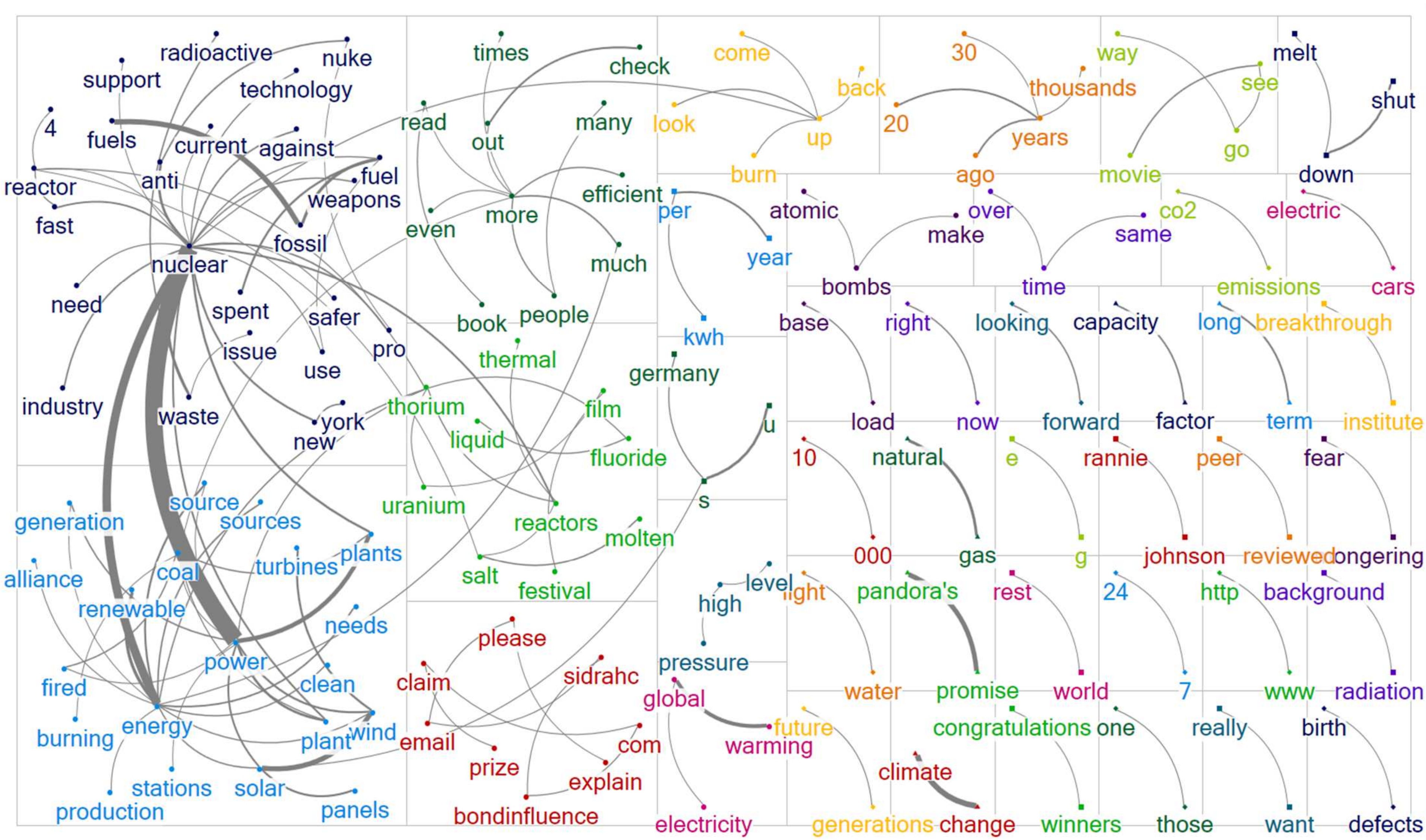


72% are male, 80 % are from U.S.
Ian Dechert: CEO at Safe Energy Association

“Public Opinion” on Social Media: Semantic Network of Posts (Stimulus)



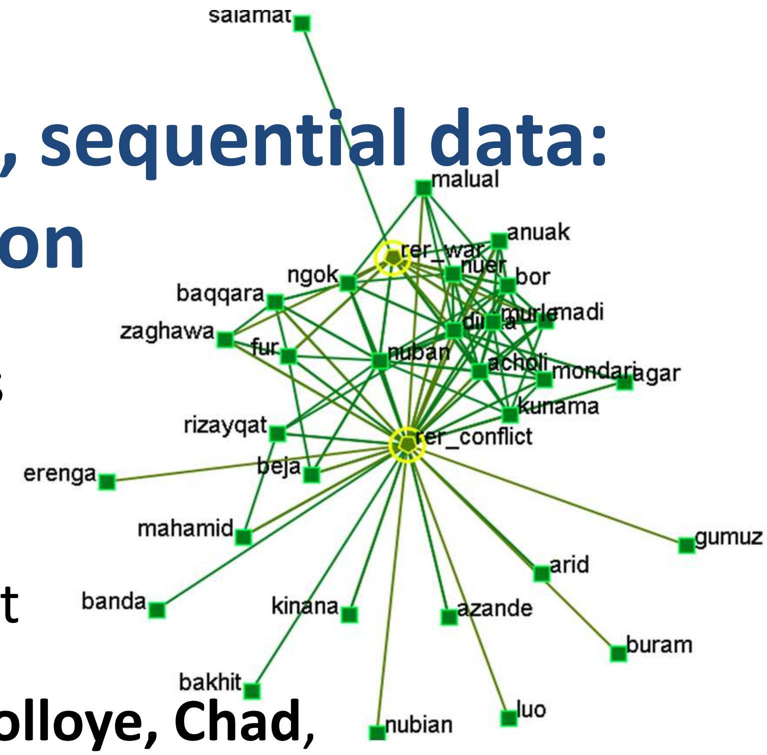
“Public Opinion” on Social Media: Semantic Network of Comments (Response)

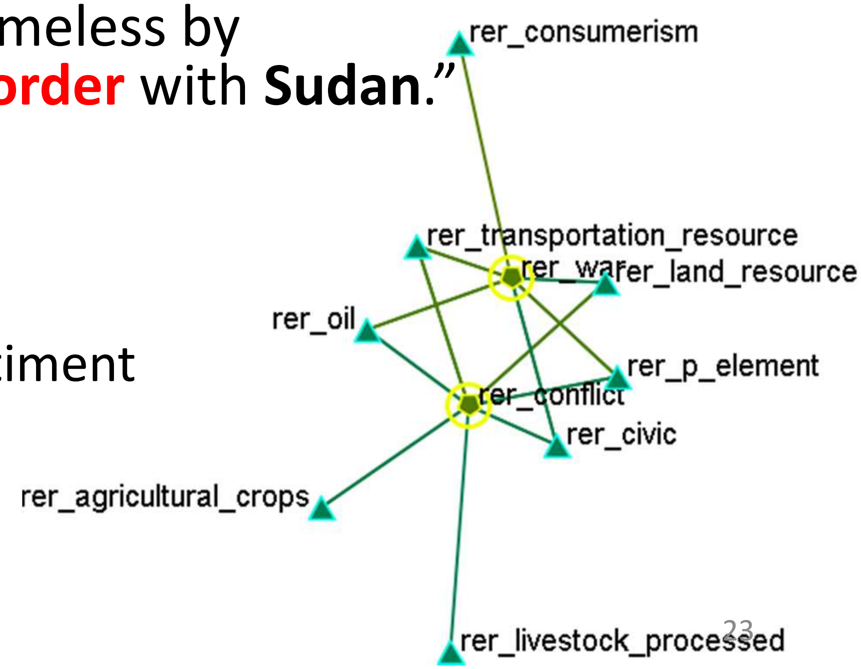


Looking under the hood: components needed

Text Mining for large, sparse, sequential data: Entity detection

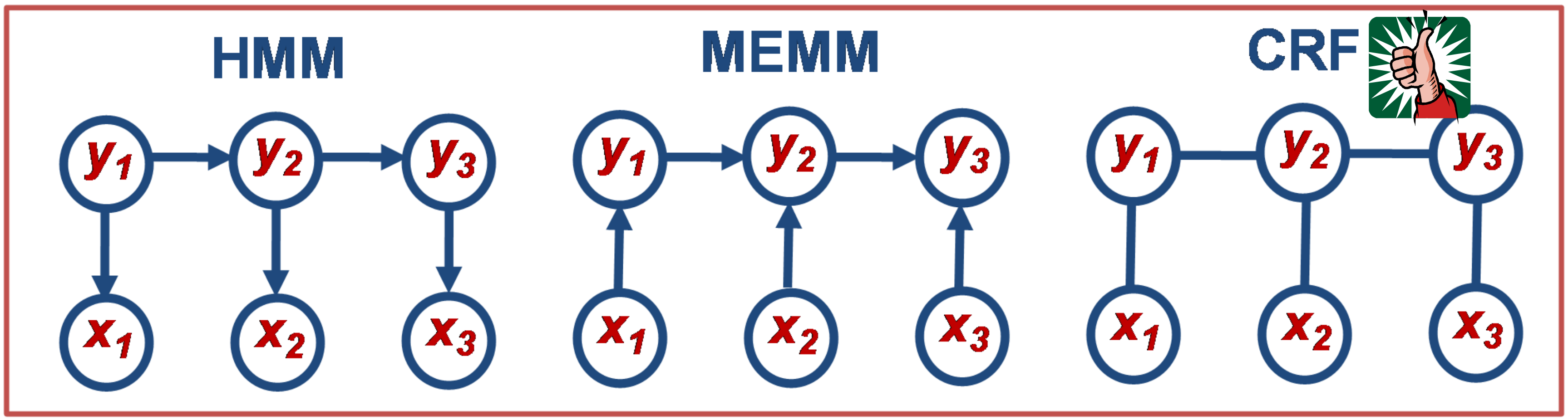
- CS: Agents, Organizations, Locations referred to be a name
- DH, CSS:
  - Also the Who? And Where?, but not referred to by a name, e.g.:
    “A **boy** and **his sister** in a **camp** in **Kolloye, Chad**, are among the **refugees** left homeless by marauding **militias** along the **border** with **Sudan**.”
  - Additional categories:
    - What? Tasks, Events
    - How? Resources
    - Why? Knowledge, Beliefs, Sentiment
    - Animals, Plants, Diseases
    - (tribal conflicts, Diesner, Carley, & Tambayong, 2012)





How to find and categorize entities in text data?

- Sequences of ***x*** (words) and ***y*** (label)
  P(x,y): generative models, e.g. Hidden Markov Model (HMM)
- P(y|x): conditional models, e.g. Maximum Entropy Markov Models (MEMM) and Conditional Random Fields (CRF)



- CRF:
  - Consider **arbitrarily large bag of features**
  - Consider and **any property of *x***, incl. long-range features

How to find and categorize nodes in text data?

- Model relationship among hidden states (y) as **Markov Random Field** (MRF) conditioned on observed data (x) (Lafferty et al. 2001)
- Compute **conditional distribution** of entity sequence *y* and observed sequence *x* as normalized product of potential functions M_i:

$$M_i(y_{i-1}, y_i \mid x) = (\exp\left(\sum_{\alpha} \underbrace{\lambda_{\alpha}}_{\text{weight}} \underbrace{f_{\alpha}(y_{i-1}, y_i, x)}_{\text{feature}} + \sum_{\beta} \underbrace{\mu_{\beta}}_{\text{weight}} \underbrace{g_{\beta}(y_i, x)}_{\text{feature}} \right)$$

$$p_{\theta}(y \mid x) = \frac{\prod_{t=1}^{n+1} M_i(y_{i-1}, y_i \mid x)}{\prod_{i=1}^{n+1} M_i(x)_{start,stop}}$$

- **Edge and transition features** plus **node and emission features**
- *f, g*: boolean feature vectors with **learned** weights

Advance in Science to Progress Digital Humanities and Computational Social Sciences

- Convex optimization over large feature space
- Tool: CRF project page, training data: BBN (LDC)
- Takes very long to train model (inference linear time)
- XSEDE allocation: parallelization on high performance computing infrastructure (factor of ten speed up on 16 processors)
- Done ☺
- In ConText

Text Mining for large, sparse, sequential data: Review Analysis

- CS: Helpfulness, Sentiment, Summarization
- DH, CSS:
  - Individual/ micro-level impact:
    - Expert (extrinsic motivation) versus laymen (intrinsic m.)
    - Nine categories:
      - (1) behavioral, (2) cognitive, (3) intentional, (4) emotional, (5) attitudinal, (6) contextualization in personal life/ personalized reflection, (7) empathy, (8) summarization, (9) ranking and mere fact of providing a review
  - Sentiment: enthusiastic versus not engaged, supportive versus non-supportive
    - Also done, in SAIL (Sentiment Analysis and Incremental Learning)

Advance in Science to Progress Digital Humanities and Computational Social Sciences

	DH as service	DH innovation
CS as service	Real boring	Digitization Data Provenance HATHI Trust (burn data)
CS innovation	Annotate data Interpretation (burn methods)	True Innovation

Acknowledgement

- This work is supported the Ford Foundation, grants 0125-6162, 0145-0558 and gifts from the Social Media Research Foundation
- Start-up allocation award from Extreme Science and Engineering Discovery Environment (XSEDE)

Thank you!
Q&A

- For questions, comments, feedback, follow-up:
Jana Diesner
Email: jdiesner@illinois.edu
Phone: (412) 519 7576
Web: http://people.lis.illinois.edu/~jdiesner