Of microscopes and telescopes in the digital humanities and computational social sciences

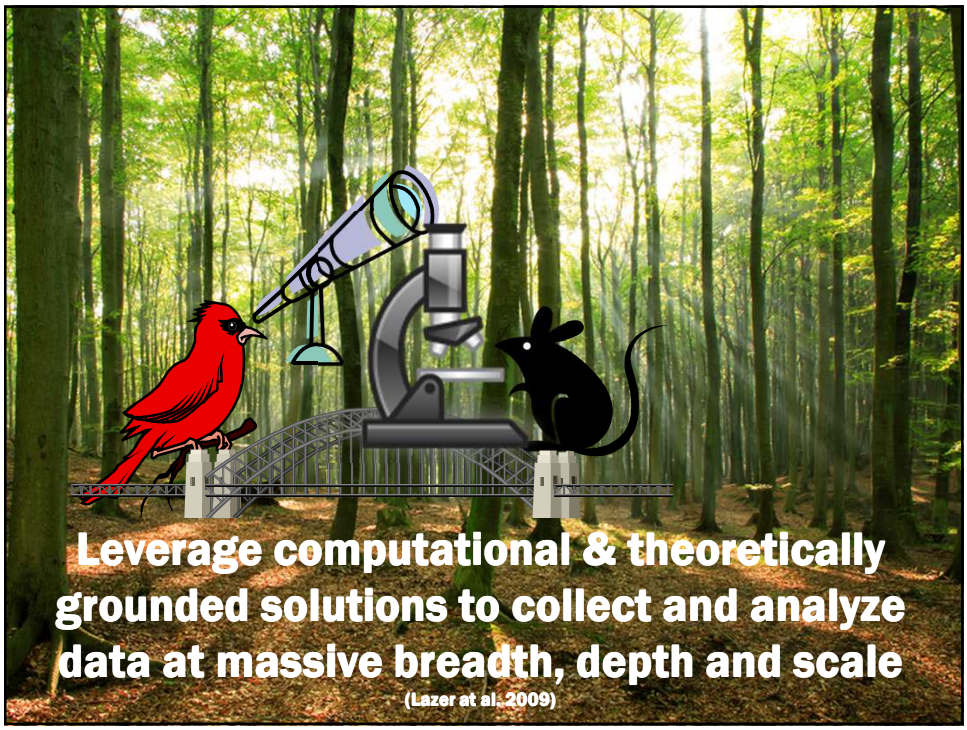
Talk at Future of Historical Network Research
Jana Diesner, PhD
Assistant Professor
The iSchool and Department of Computer Science
University of Illinois Urbana-Champaign

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

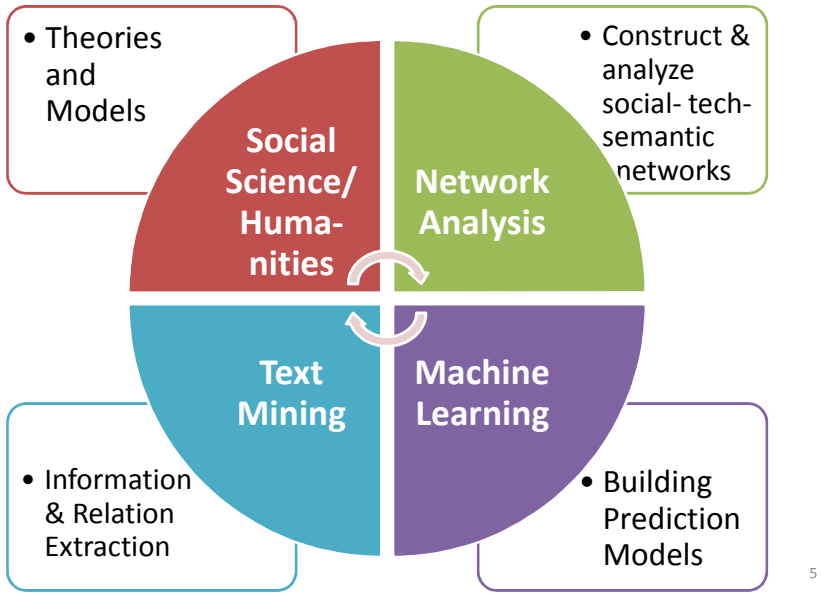
GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

1


Data, methods, tools for building theories in social sciences and humanities
Traces of digitized, archival, contextualized behavioral data (structure and content)


Leverage computational & theoretically grounded solutions to collect and analyze data at massive breadth, depth and scale
(Lazer et al. 2009)

Example: My Approach: Words and Networks



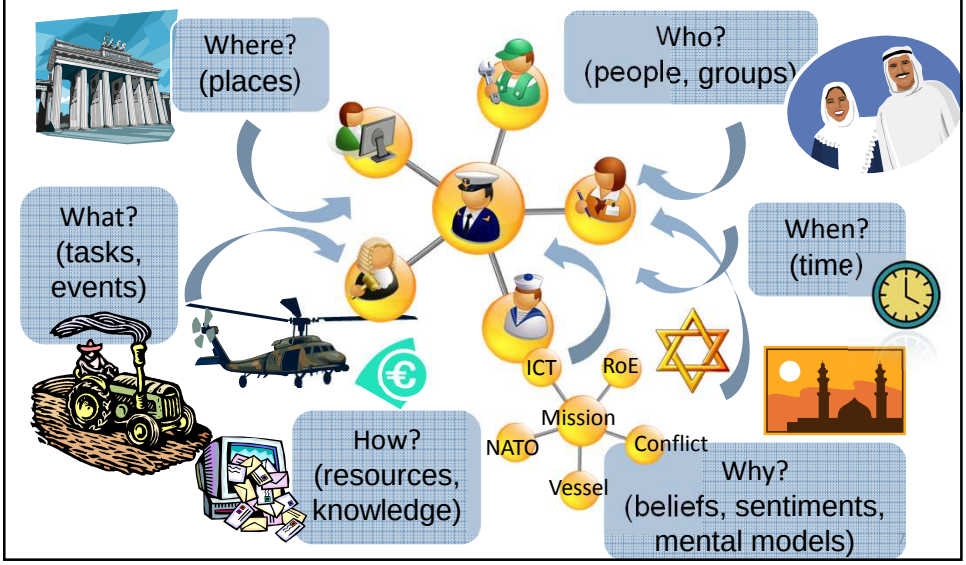
5

From Words to Networks: Example: Sudan

- **Task:** Develop, evaluate, apply method and technology for extracting socio-technical network data from large-scale text corpora to answer questions about the Sudan.

Diesner J, Tamabyong L, Carley KM (2012) Mapping socio-cultural networks of Sudan from open-source, large-scale text data. Computational and Mathematical Organization Theory (CMOT)

From Words to Networks: What node classes to consider?



How to find and categorize nodes in text data? Basic recipe for probabilistic solution

- Get some labeled ground-truth data (BBN)
- Build a **classifier**/model (h) that **for every sequence of words (x)** and label per word (y) **predicts one category per word** (y = h (x)), incl. for new and unseen text data
- Exploit clues from text data (lexical, syntactic, statistical)
- Train and validate the model
- Get good accuracy (compare to intercoder reliability) (we made model available in end-user product AutoMap)
- Apply prediction model to text data (~ 80,000 files)
- Link nodes (e.g. based on co-occurrence, proximity)
- Network data! Analysis!

Diesner J, Carley KM (2008) Conditional Random Fields for Entity Extraction and Ontological Text Coding. Journal of Computational and Mathematical Organization Theory (CMOT), 14(3), 248 – 262.

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:

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

(weight, feature; weight, feature)

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

- **Edge and transition features** plus **node and emission features**
- *f, g*: boolean feature vectors with **learned** weights
- Tool: CRF project page, training data: BBN

How good is it?

Class model	Class	Specificity	Subtype	Example	States, Edges	Time (300)
1	x			agent	11, 121	17.5 hours
2	x	x		agent, specific	16, 256	1.25 days
3	x		x	agent, political	32, 1024	3.1 days
4	x	x	x	agent, spec., pol.	45, 2025	5 days

Variable	Accuracy	Training Time
Baseline	large	small
Syntax features (POS)	small	small
Lexical features (dict, hard match)	large	small
Iteration rate	large	large
Complexity of class model	small	large

	Boundary Model	Class model 1 (cat.)	Class model 2 (cat. + spec.)	Class model 3 (cat.+ sub.)	Class model 4 (cat+spec+sub)
Precision	93.2%	91.4%	91.9%	90.4%	90.8%
Recall	92.5%	89.7%	90.0%	88.6%	88.9%
F	92.9%	90.6%	90.9%	89.5%	89.8%
Bound. & Class (rules)		**Model 1**	**Model 2**	**Model 3**	**Model 4**
Precision	n.a.	89.7%	90.0%	88.6%	88.9%
Recall		87.7%	87.7%	86.4%	86.5%
F		**88.7%**	**88.8%**	**87.5%**	**87.7%**

Network Data! Analysis!

Activity:

Degree Centrality	03	04	05	06	07	08	09	10
Omar al-Bashir	3	3	2	1	1	1	1	1
Ali Osman Taha	1	2	3	4	3	3	3	3
John Garang	2	1	1	3	3	4	6	8
Salva Kiir Mayardit	8	10	4	2	2	2	2	2
Hosni Mubarak	4	7	5	6	9	8	4	6
Sadiq al-Mahdi	6	5	10	9	5	7	8	4
Hassan al-Turabi	5	6	7	10	5	8	9	5
Abdul Wahid al Nur	10	9	9	8	7	4	5	7
Yoweri Museveni	7	8	7	6	11	10	7	8
Kofi Annan	9	4	6	5	8	11	11	11
Deng Alor	11	11	11	11	10	6	9	8

Control:

Betweenness Centr.	03	04	05	06	07	08	09	10
Omar al-Bashir	1	1	1	1	1	1	1	1
Salva Kiir Mayardit	6	10	2	5	2	2	2	2
Ali Osman Taha	4	3	3	7	6	7	5	4
John Garang	3	6	5	4	4	6	7	7
Sadiq al-Mahdi	2	8	10	2	7	5	6	3
Abdul Wahid al Nur	8	4	7	8	3	4	3	6
Kofi Annan	7	2	4	3	10	11	8	10
Yoweri Museveni	5	5	9	6	5	9	8	10
Deng Alor	8	10	10	9	9	3	8	5
Hosni Mubarak	8	9	8	11	8	8	4	8
Hassan al-Turabi	8	7	6	10	11	10	8	9

Close to power:

Eigenvector Centr.	03	04	05	06	07	08	09	10
Ali Osman Taha	1	2	3	3	3	3	3	4
Omar al-Bashir	3	3	5	2	2	2	2	3
Salva Kiir Mayardit	7	11	4	1	1	1	1	1
John Garang	2	1	1	4	4	4	7	6
Hosni Mubarak	4	5	6	5	11	5	4	7
Kofi Annan	8	4	7	6	6	11	11	1
Yoweri Museveni	9	8	8	7	9	6	5	8
Hassan al-Turabi	5	7	10	8	8	10	8	5
Sadiq al-Mahdi	6	6	9	9	7	8	10	6
Deng Alor	11	1	11	10	5	7	9	10
Abdul Wahid al Nur	10	9	11	11	11	10	6	9

Triads

Triads	03	04	05	06	07	08	09	10
Omar al-Bashir	1	1	1	1	1	1	1	1
Ali Osman Taha	2	3	3	4	4	3	2	2
John Garang	3	2	2	2	2	6	7	7
Salva Kiir Mayardit	7	10	4	3	3	2	3	3
Hosni Mubarak	7	4	5	6	6	8	4	5
Sadiq al-Mahdi	4	7	7	7	6	7	7	3
Abdul Wahid al Nur	10	9	9	7	4	5	5	7
Kofi Annan	7	5	5	5	11	11	7	7
Yoweri Museveni	6	6	8	9	9	10	6	5
Hassan al-Turabi	5	8	9	9	8	9	7	7
Deng Alor	10	10	9	9	10	4	7	7

- President North: Known performer
- President South: Now established
- Legacy of religious leaders
- Presence of neighboring presidents

2003	2004	2005	2007	2010
Darfur conflict	Continuous civil war (since 1993)	Comprehensive Peace Agreement Garang 1st VP, followed by Kiir Autonomous South Sudan	SPLA withdraws from government	Votum in South Sudan about Separation

Sudan: Results (Groups)

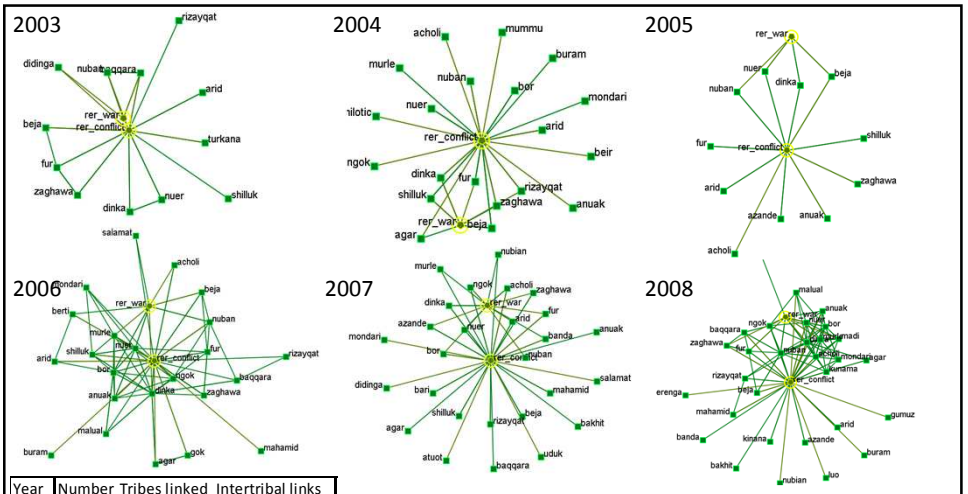
Degree Centrality	03	04	05	06	07	08	09	10
United Nations	4	2	1	1	1	1	1	5
Rebel Groups	1	1	2	3	4	3	2	3
Military	2	3	3	2	2	2	4	2
SPLA	11	6	5	4	3	4	3	1
Security Council	5	5	4	5	5	5	5	6
Sudan government	3	4	6	6	8	8	9	7
Nat. Congress Party	6	9	9	8	6	7	10	4
African Union	8	7	8	7	7	9	7	9
Inter. Criminal Court	9	11	7	11	9	6	6	9
Dinka	9	10	11	9	10	10	8	8
Churches	7	8	10	10	11	11	11	11

Betweenness Centr.	03	04	05	06	07	08	09	10
Military	1	1	3	3	1	1	2	1
United Nations	3	6	2	2	3	2	1	3
SPLA	11	3	1	1	2	3	5	2
Rebel Groups	4	2	4	4	7	5	3	4
Sudan government	2	4	5	8	4	7	6	10
Nat. Congress Party	6	9	8	5	5	4	8	7
Churches	5	7	9	10	6	6	9	9
Dinka	8	5	6	6	8	11	11	6
African Union	7	8	7	11	10	10	10	5
Inter. Criminal Court	11	11	10	9	9	8	4	11
Security Council	9	10	11	7	11	9	7	8

Eigenvector Centr.	03	04	05	06	07	08	09	10
United Nations	4	2	1	2	1	2	1	5
Military	2	3	3	1	2	1	5	2
Rebel Groups	1	1	4	3	4	3	6	3
Security Council	5	5	2	4	5	4	2	8
SPLA	11	6	5	5	3	5	7	1
Sudan government	3	4	7	6	8	7	8	6
African Union	8	7	8	7	6	9	4	10
Inter. Criminal Court	11	10	6	9	9	6	3	7
Nat. Congress Party	6	9	10	8	7	8	9	4
Churches	7	8	9	10	10	10	10	11
Dinka	9	11	11	11	11	11	11	9

Triads	03	04	05	06	07	08	09	10
Military	1	1	1	1	2	1	6	1
United Nations	4	3	2	2	1	4	1	2
Rebel Groups	2	2	4	4	4	2	4	5
SPLA	11	5	3	3	3	3	2	4
Sudan government	3	4	5	7	5	7	4	6
Nat. Congress Party	5	9	10	8	6	6	9	3
African Union	8	6	6	6	7	10	7	9
Security Council	7	7	7	5	8	9	8	8
Inter. Criminal Court	11	11	8	9	11	5	3	7
Churches	6	8	9	10	9	8	10	11
Dinka	9	10	11	11	11	11	11	10

- Strong presence of armed forces
- Strong influence of external groups
- Within top 10 Sudanese groups:
  - Dinka, Nuer (ethnic groups/ tribes)



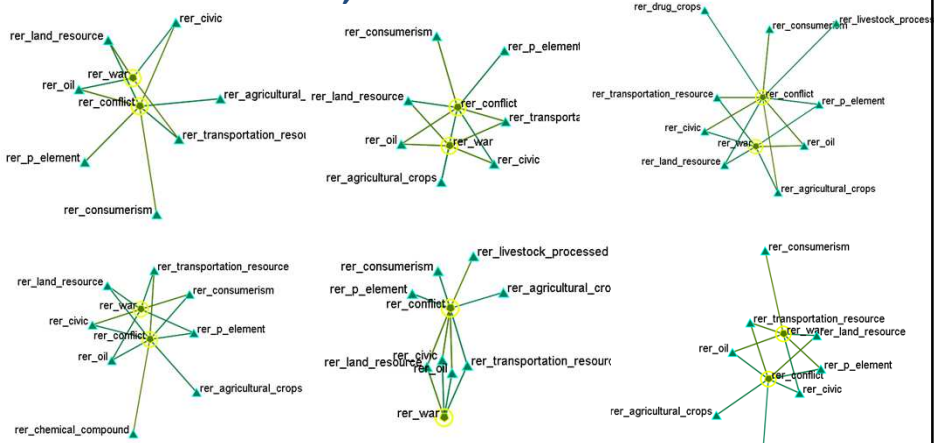
Year	Number of tribes	Tribes linked to conflict or war	Intertribal links for pairs linked to conflict or war
2003	32	38%	32%
2004	44	45%	66%
2005	33	39%	40%
2006	46	50%	83%
2007	47	62%	78%
2008	50	60%	65%
2009	28	68%	95%
2010	27	56%	100%

- High and increasing rate of tribes associated with conflict or war

Sudan: Results

13

Sudan: Results: Conflict, War and Resources



- Conflict: Agriculture, Livestock (farmers vs. herders)
- War: Land Resource (concept of *dar*)
- Conflict and War: Oil, Civic, Transportation

14

Example: Assessment of Impact of Social Justice Documentaries

- **Storytelling**: documentaries create memories and imagination, stimulate sharing (Rose 2012)
- **Impact**: documentaries aim to cause **change** in people's knowledge and/or behavior (Barrett & Leddy 2008)
- **Status quo:** funders, practitioners, scientists agree on **strong need for rigorous impact assessment**, but amount and depth of **prior reports and work limited**
- **Our questions**: How can we know if a documentary has impact? At what point in the life cycle of a film can we answer this question?

15

Our Solution

- **Solution**: develop, implement, apply, evaluate computational solution for gaining **broad** (quantitative) **and deep** (qualitative) understanding of impact (change)
- **Assumption**: films produced, screened, watched as part of large and **continuously changing ecosystems of stakeholders and themes**
- **ConText**: track, map, analyze structure, functioning and dynamics of **web of stakeholders and themes** associated with main issue addressed in a movie at scale
  - **Look no hands – empirical data analytics: network analysis** (map and analyze social networks) paired with **natural language processing/ text mining** (semantic networks, salient concepts and topics, valence of content)
  - **This is no fishing expedition – we have a theory**: developed CoMTI (content, medium, target, impact), a comprehensive measurement framework

Our Solution: Research and Development Process

Theory	• Comprehensive lit review -> develop framework of relevant dimensions/indicators of media impact
Operationalization	• Translate indicators into metrics and indices
Methods & Algorithms	• Map indices to methods and algorithms suitable for analyzing large-scale, empirical data
Data Collection	• Empirical: news coverage, social media, focus groups, ….
Technology	• ConText: Comprehensive tech review -> build tool from open source code and from scratch
Analysis & Interpretation	• Apply technology to various data sources on various movies
Evaluation	• Assess accuracy and performance of methodology and technology

Our Methodology in a Nutshell

Understand problem space: (Where) is impact possible?

- Map public discourse and key players (different types of influence and power) related to main theme(s) of film prior to release (baseline model)
- Collect, analyze, fuse data from news, social media, focus groups
- Map opportunity space for linking up to people and themes, and strategic resource allocation (mobilize (social) capital)
- Identify unpromising topics early (no or uncontroversial discourse)

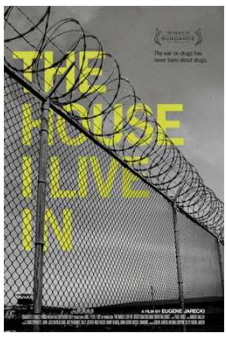
Understand film's message: Impact w.r.t. what?

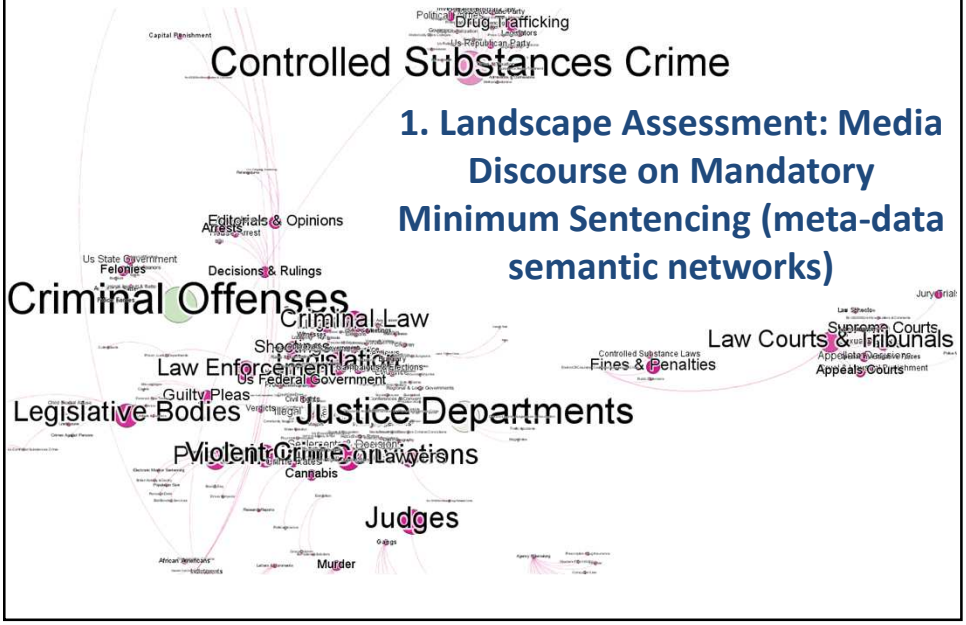
- Analyze transcripts (ground truth model)

Understand film's impact: Has the needle moved?

- Reassess map of discourse and key players: New links, new themes that connect movie to baseline? Baseline connected or closer to ground truth?

ConText in Action: The House I Live In (Eugene Jarecki 2012)



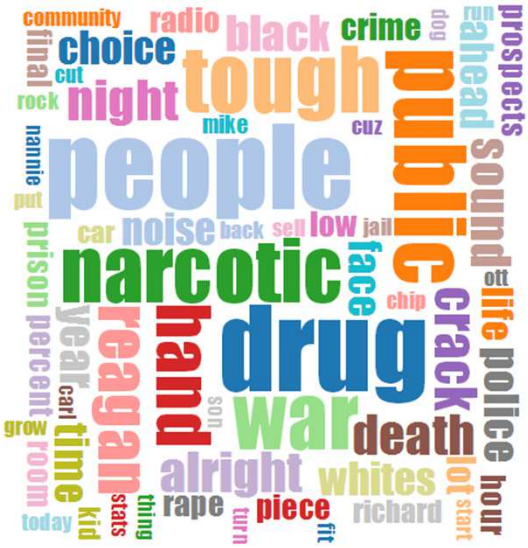


1. Landscape Assessment: Media Discourse on Mandatory Minimum Sentencing (meta-data semantic networks)

1. Landscape Assessment: Media Discourse on Mandatory Minimum Sentencing (content topic modeling)

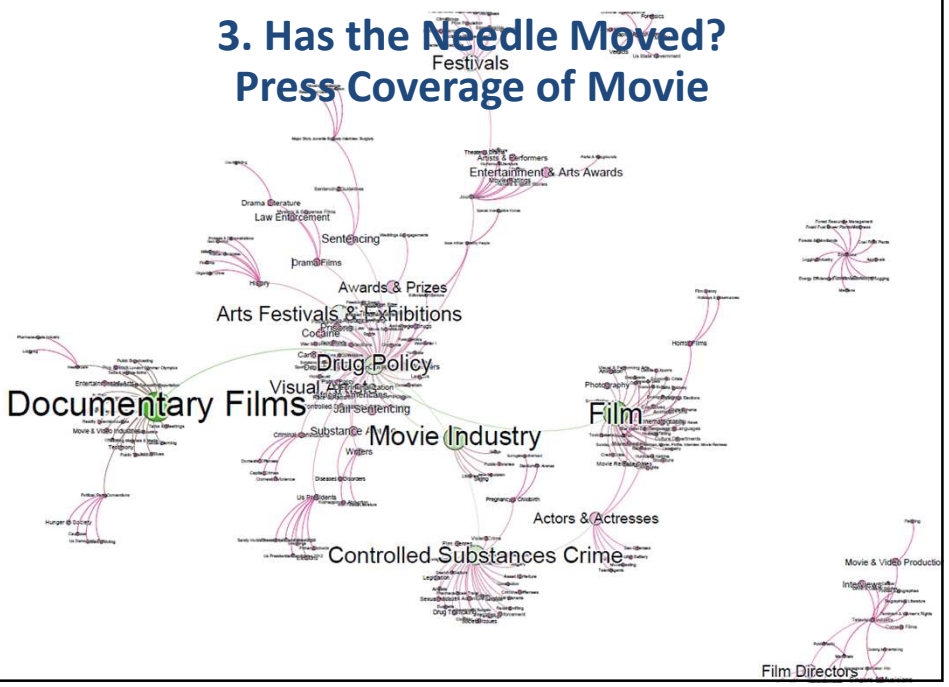
framed as social issue (people in center of public discourse)

2. What's in the Movie? (topic modeling of transcript)

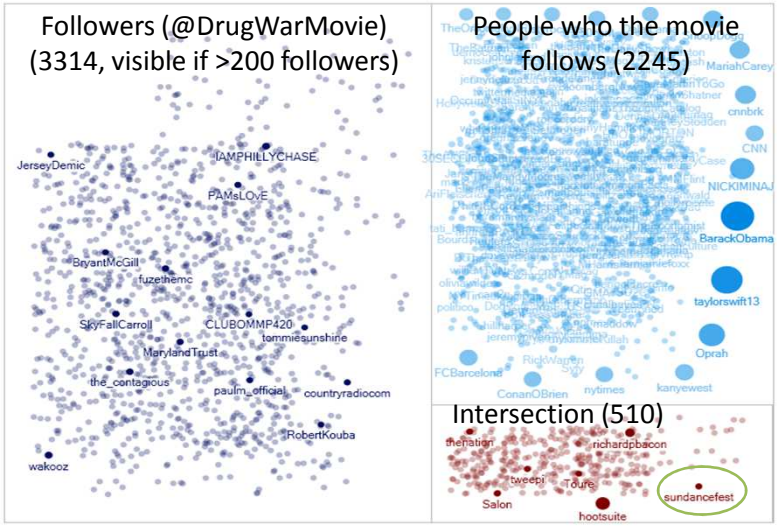


large common denominator: people (social issue), difference: media: prisons and violence, movie: drugs and related politics

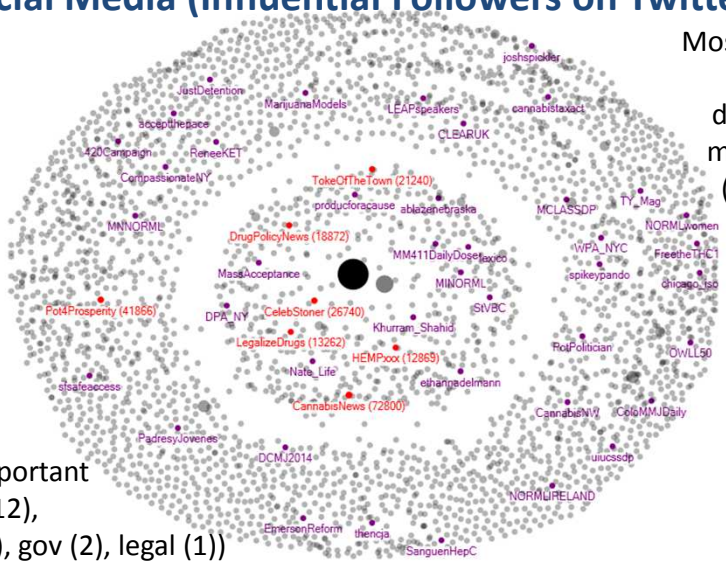
3. Has the Needle Moved?
Press Coverage of Movie



3. Has the Needle Moved?
Social Media (Twitter network of followers)



3. Has the Needle Moved?
Social Media (Influential Followers on Twitter)

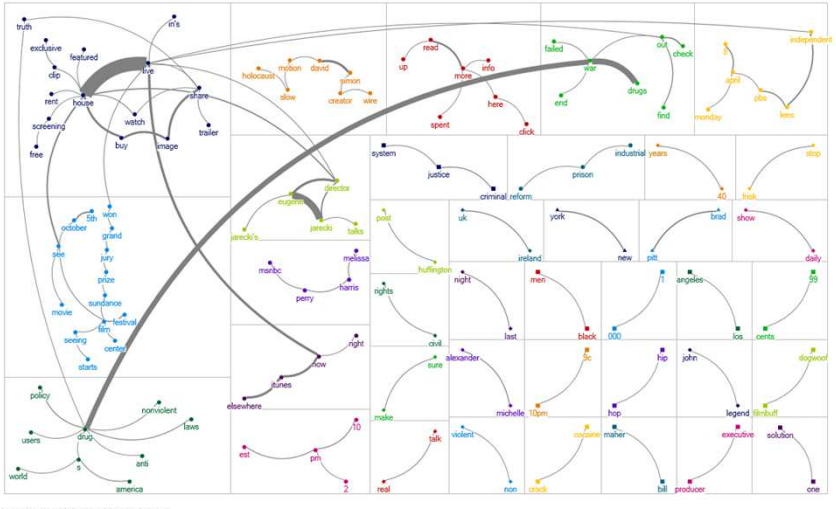


Most other: writers, directors, musicians (comfort zone)

Red = important (media (12), NGO (33), gov (2), legal (1))

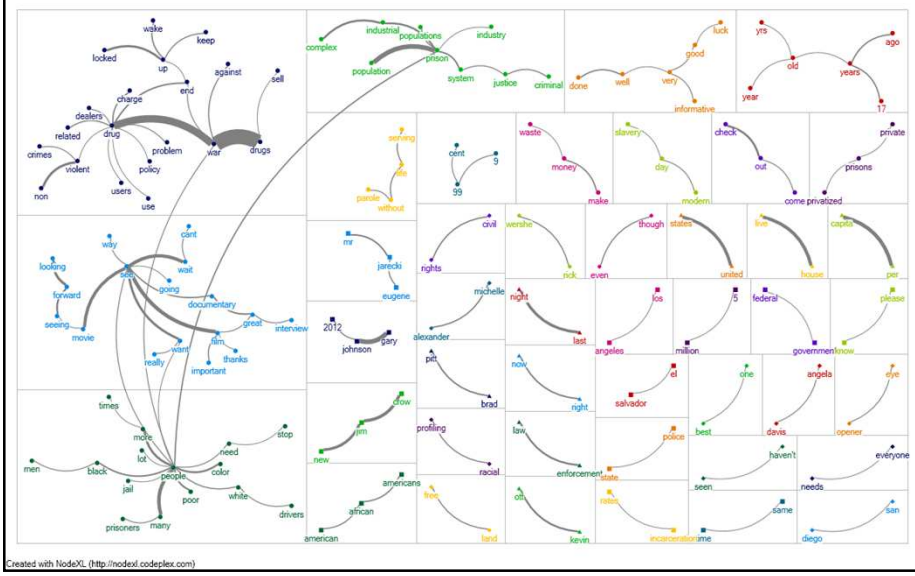
good spot for building up your social capital here (8/48 >100K followers)

3. Has the Needle Moved? Social Media Campaign (Facebook Fanpage Posts, 428)



Created with NodeXL (http://nodexl.codeplex.com)

3. Has the Needle Moved? Social Media Reactions (Facebook Comments, 1412)



Acknowledgements

- Impact Assessment: FORD Foundation, grant 0125-6162.
- Sudan: National Science Foundation (NSF) IGERT 9972762, the Army Research Institute (ARI) W91WAW07C0063, the Army Research Laboratory (ARL/CTA) DAAD19-01- 2-0009, the Air Force Office of Scientific Research (AFOSR) MURI FA9550-05-1-0388, the Office of Naval Research (ONR) MURI N00014-08-11186, and a Siebel Scholarship.

References

- Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., . . . Gutmann, M. (2009). Life in the network: the coming age of computational social science. *Science, 323*(5915), 721-723.
- On movie impact assessment:
  - Diesner J, Aleyasen A, Kim J, Mishra S, Soltani S (2013) Using Socio-Semantic Network Analysis for Assessing the Impact of Documentaries. WIN (Workshop on Information in Networks) 2013, New York, NY
- On Sudan project:
  - Diesner J, Carley KM, Tamabyong L (2012) Mapping socio-cultural networks of Sudan from open-source, large-scale text data. Computational and Mathematical Organization Theory (CMOT) 18(3), 328-339.
  - Diesner J (2012) Uncovering and Managing the Impact of Methodological Choices for the Computational Construction of Socio-Technical Networks from Texts. Technical Report CMU-ISR-12-101.

Thank you!

Q&A

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