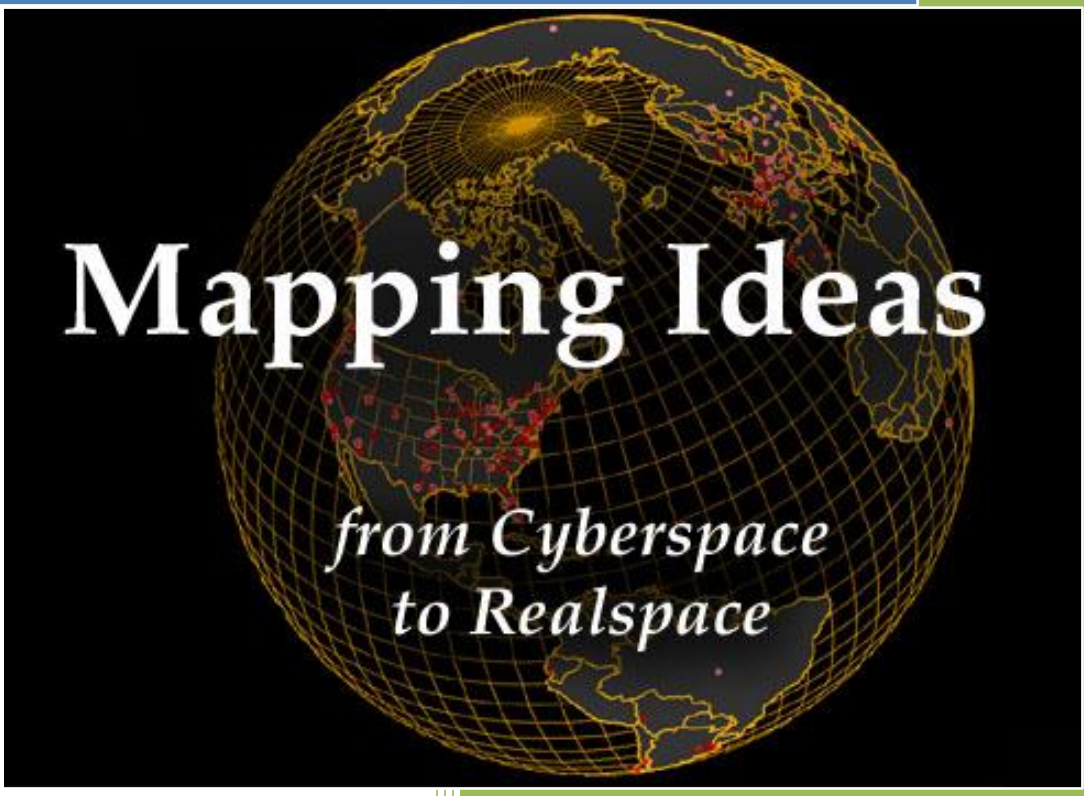
2012

CDI Specialist Meeting Report



Editors: Ming-Hsiang Tsou, Sarah Wandersee, Jiue-An Yang, Daniel Lusher.
(Mapping Cyberspace project), SDSU

Funding: NSF-CNS # 1028177
8/1/2012 – 8/02/2012

Table of Contents

EXECUTIVE SUMMARY

This specialist meeting (workshop) is funded by this NSF-CDI project (#1028177), ***Mapping Cyberspace to Realspace.*** The goal of this workshop is to foster the multidisciplinary collaboration in related research disciplines, including geography, linguistics, computer science, political science, and communication. The two-day workshop (August 1 and August 2, 2012), organized by San Diego State University, brought together 13 specialists drawn from the many disciplines with interest in these issues. The workshop assessed the current state of the art, identify and prioritize a research agenda, and begin the development of a research community of collaborating scholars working on these issues. The meeting included plenary presentations by invited experts, and ample time for small-group discussion of the issues. This workshop generated a final report published on the project website (http://mappingideas.sdsu.edu/?page_id=403).

Specific research questions were addressed in the workshop, including:

1. How to quantify, map, and understanding idea diffusion over space and time?
2. How can various spatial relationships and social impacts between cyberspace activities and real-world events be explained?
3. To what extent and in what ways can social network diffusion and internet diffusion be integrated geospatially?
4. What are the shared features of the problems of the sentiment analysis and group/political-orientation identification and what are the best domain-independent approaches to both, given the current state of the art?
5. How to validate these cyberspace information and explanations?

During the workshop, our group discussion identify the following important research topics for our scientific research community:

- The Veracity of Geolocations in Cyberspace,
- Sentiment Analysis for 2012 Presidential Election,
- Collective Identity,
- Benefits and Challenges of Working with Social Media, Twitter
- Geolocation Implications,
- Communication Theory,
- Text/Data Mining & Theory Building, and
- Space-Time Modeling

This report also highlights the plenary suggestions for the next steps of this CDI project, including short-term actions and long-term strategies, with additional resources and references available related to these research topics.

KEY IDEAS

Benefits and Challenges of Working with Social Media, Twitter

Twitter owns more than 140 million users and generates over 340 million tweets in a day which makes information from such social media massive and become a set of Big Data. However, questions remain towards the credibility of these social media data. Taking tweets as example, are all tweets generated by real users? Or more by those tweeting-robot which well known as water-army? Furthermore, even if all tweets are generate by real users, it is still questionable to draw direct connections between a massage and the user's real thought on the topic. In the sense of social science, *are human behaviors in social networks being emulated?*

The *user* of social media is another aspect to be concerned. Social media driven by the advances of Information and communications technology (ICT) obviously have the inherent issue of digital dived. User of social media only represents the population that has the ability to afford devices and connectivity to the linked social network. We also have limited understanding towards *who* are the users of social media. The word of *who* here is beyond the basic demographics (age, gender …etc.) but at the level of self-identity, group affiliations, and in a sense of community. For specific topics that are trending on the social media, what motivates people to join the discussion and what groups are they involved with? What are the mechanisms to drive the diffusion of ideas in social networks? Studies on contemporary social network are needed to answer these questions and bring more confidences to utilize data from social media in understanding our Cyberspace.

Geolocation Implications

Twitter can be attractive to geographers because each tweet contains locational information. There are two types of location information associated with each Tweet on Twitter: a geo-tagged location when the user enables GPS on the device, or a self-reported location that is specified in the user profile. From previous studies the overall percentage of including a geo-tagged location (GPS) a tweet is 1% while other 99% remains as self-declare locations. We have seen most Twitter researches utilizing the 1 % of geo-tagged location trying to understand social events and human behaviors. However, is information from the 1% generalizable of the rest 99% or the population of our interest? The opposite group argues that by looking at colossal amount of 1% tweets, issue of geo-tagged tweets and self-declare may not be significant. There are gray zones that require more efforts on the representative of the 1% geo-tagged tweets. This also leads to the question that if it is possible and needed to collect all the tweets to have better understanding towards behaviors in social media. Except of social media, we also concern the methodology on assigning location to websites. Using IP address to retrieve real location has many uncertainties and domain name could be the next target to be validated.

Communication Theory

We ask the question if there is a need to build new communication theory which can summarize our findings on Twitter, expertise in our group tackled the topic from many aspects. Building networks from the massive data could be a starting point to visualize the Cyberspace and by calculating centrality metrics we see the differences of characteristics between networks. Studying nodes that serve as the bridge of connecting clusters is the second step. We learn the mechanism of how innovation diffuses through social network by reviewing the influential power of opinion leaders. The quick conclusion is there is the need to form a new communication theory that integrates the ontology model because the patterns of communications largely vary by the topics and types of events. Methodology will not be limited to network analysis and can be combined with linguistic meanings, gamification, and spatial analysis.

Text/Data Mining & Theory Building

We see content analysis and text mining as fundamental challenges when trying to understand Cyberspace. Filtering and post-processing data is important when we don't have the luxury to archive everything from social media. Two approaches have long been welcomed in social science: the theory-based approach trying to prove what we thought the world should like, and the data-driven approach to discovery patterns then build theories from our daily life. With big data from social media we need to ask questions about data before we start studying it and trying to understand it. This is where the theory-leading approach comes in and why we should encourage more model and theory building. Models should be test with data then be adjusted or discard. Theory will have significant relation to data especially in explanatory and predictive reasons. Thus, social science is considered to be the bridge of massive social media data and new theories to be built.

Space-Time Modeling

There are various ways to approach space-time models, with methods and theory still being developed. We need to find a set of metrics that describe the space-time changes while also allowing for the complex dynamics of issues within different domains (e.g., epidemiology, conservation, politics). In considering these models, we may not find a one-size-fits-all approach or method. For example, kernel density is useful for ends like visualizing hotspots, but it should not be the only method used. In addition, p-values may not be meaningful in all cases. Throughout contemplation and analysis of these topics, we need to conceptualize meanings of results and the applications of our findings. We must be creative in discussing and developing approaches for space-time analysis. It is also vital that we venture into making predictions. We will often be wrong, but we will learn from those results and be able to apply the findings in useful ways.

NEXT STEPS

Our specialists identified a few short term actions (within one year) and long term strategies (3-4 years) for related web content search and analysis.

SHORT-TERM

1. Attempt to understand the overall communication properties of social media, Twitter in particular
2. Look for domain name databases
3. Break into focus groups to debate each issue
4. Dynamic model- not kernel density static model
5. Focus on a specific dataset and then determine the best approach for it
6. Test tweet data with LIWC
7. Construct network graph from social media data to find strong ties and opinion leaders

LONG-TERM

A. Look for other research on the demographics of the 1% of geotagged tweets
B. Use simple diffusion models then determine how well these fit data and then adapt
C. Compare the content analysis results from LIWC and Mark's other linguistics methods
D. Building prediction models from data for future events

Discussion Points I: Project Strengths

- Valid knowledge is being discovered within this project-
- Such diversity provides different aspects from different projects.
- The participants provided excellent presentations and discussions.
- The meeting presented a great range of thought and approaches to working with data. There is real theoretical work being done.
- Already gathering large numbers of geotagged tweets
- The technological abilities of members of the project to create and display complex patterns
- Focused on tracking these space time issues
- The meeting presented a great range of thought and approaches to working with data. There is real theoretical work being done.
- Large tweet collection already
- Domain expertise from communication (Spitzberg) and social science (Gupta)

Discussion Points II: Issues/Challenges

- Determining tweet sentiment is difficult
- Generalizeability of twitter community
- Brian believes we have been theory rich and data poor.
- Collecting data is expensive and challenging.
- Challenges: psychology communication. The person has the motive, cultural experience, generate messages for certain people.
- We have lots of challenging data. Challenge with our disciplines is disembodied voice. Unit of explanation is the individual (person generate messages for reasons) now messages are only understood in batches. Want to explain batches collectively, but where is the person in the explanation?
- How to connect what people are doing with their messages and how it can be traced back.
- Challenges in field are contextual. There are some trends to look for such as gaps, break points, and lags.
- With regard to twitter goes back to problem of (we need to know who, what, where when to learn something from the daily source.
- Need to go beyond and see network. Need to know what sticks and persuades how to help people who are influenced by misinformation.
- For the future, it would be useful if many groups set up a collaboration infrastructure.
- There are a lot of different ideas being presented in the project. Text mining is a fundamental challenge in all this. We tend to focus a lot on diffusion and the network. All these questions boil down to can we mine the info from the text we want.
- IP address geolocation is inaccurate
- There is little data on how representative the 1% of geotagged tweets are in the overall population
- Few data is geolocated
- Sharing data
- Old metrics are not optimal in describing the project data
- Currently, the project does not have a theoretical framework to go with the data and tools
- Should the project focus more on explanation or predication
- Do not have a quantitative spatial index to measure change
- Spatial data is never random- p-value test is not necessarily helpful
- Integration of different indicies/measures
- Need for a model and statistical analysis methods
- Where can we publish?
- Who are experts? experts from various fields or experts in GIS with knowledge of other fields?
- Uncertainty from tweet contents, how to correctly interpret the emotion from word use
- Challenges of location

- Issue of each keyword having different dynamics and different sets of issues.
- From large amount of data, what to look at and how to explain what we find
- Difficult to do sentiment analysis of Twitter data

Discussion Points III: Suggestions

- It would be interesting to map pro and con (for vaccines)
- May want to look for collaboration within this group (to create social networks)
- Just need the overall theory of communication- emphasis on Diffusion of Innovation
- Find the element of gamification within our own research
- There should be collaboration among the participants. For instance, involvement through the Internet or other means such as discussion forums under the project's website.
- Twitter data presents an opportunity for some collaboration among the participants if possible.
- The project team could come up with some way to involve the participants to keep the momentum going.
- Collectively as a group, it should be considered publishing what we know, this are the tools using to navigate.
- For the future, it would be useful if many groups set up a collaboration infrastructure.
- Adapting a multi-disciplinary approach.
- For future research: devices of distortion; more changes to include network analysis approaches.
- Looking at the message, demographics, and more importantly natural structures. What other sources could be useful for research beside twitter data?
- Examine domain names as a potential method of website geolocation
- Find out if geotagged tweets are in any way different than overall population
- Examine potential non-Euclidian space frameworks for the patterns
- Use amazon's mechanical turk for labeling tweets
- Use a focus group to explore these issues
- Define exact questions specific to project- hard to create unified model
- Open meetings/conferences to more people
- Possibly new journal
- Need a way for the group to share resources, for now will use email to connect
- Collaboration book
- There should be collaboration among the participants.
- Implementation of LIWC on sentiment analysis with project's tweet data
- GIScience has lots of discussion on geographic names/gazetteers that could provide some ideas on aggregation
- Work with sets of data instead of two dataset comparisons at one time
- Building network from users and linkage
- Network analysis to find special features from the large network

- Not only build theory but make prediction models
- Provide options for making collaboration easy for the future: perhaps email list, Google group, Twitter, teleconference, etc…

Resources & Further References

- Tools:
  - http://newsmap.jp/
  - Amazon's Mechanical Turk
  - www.trendinghashtags.com
  - trendsmap.com
  - researchblogging,org
  - academia.org
  - academiamap.com
  - http://memetracker.org/
  - GeoDa
- Data centers/Contacts:
  - University of Arizona, Artificial Intelligence Lab, extremist websites archive (10 years of data)
  - http://newseum.org/
  - Library of Congress
- Publications/Sources:
  - http://www.ncbi.nlm.nih.gov/pubmed/
  - Book: False Paradigm
  - Remote Sensing may have patterns that can be used
  - Literature in GIS since 2004 on event-based modeling
  - MITRE's social radar report on the usage experience of LIWC with tweets
  - Malcolm Gladwell's Tipping Point
  - Watts and Dodds Diffusion of Innovation
- Other Projects:
  - Geovisualization – Terri Schiavo case (student project, 2005)
  - WordNet

Report Prepared by:

Ming-Hsiang Tsou (PI), Sarah Wandersee (Graduate Assistant, Ph.D. student), Jiue-An Yang (Graduate Assistant, Ph.D. student), Daniel Lusher (Graduate Assistant, Masters student).

PARTICIPANTS

National Visiting Committee (NVC)

Photo	*First Name/ Last Name*	*Telephone/ Email*	*Title*	*Organization*	*Address*
	Shih-Lung Shaw	sshaw@utk.edu	Professor	Department of Geography at the University of Tennessee	The University of Tennessee 304 Burchfiel Geography Building Knoxville, TN 37996-0925
	Edna Reid		Analyst	Department of Justice	
	Clayton Fink	finkcr1@jhuapl.edu	Senior Software Engineer	Johns Hopkins University, Applied Physics Laboratory	11100 Johns Hopkins Road, Laurel, Maryland 20723

Specialists

	Michael Thomas	Mthomas304@att.net Michael.thomas3@eucom.mil	Lieutenant Colonel	United States Air Force, Penn State University	Department of Geography 302 Walker Building The Pennsylvania State University University Park, PA 16802
	Jennifer Mathieu	jmathieu@mitre.org	Systems Engineer	The MITRE Corporation	
	Les Servi	lservi@mitre.org	Group Leader	The MITRE Corporation	MITRE-Bedford 202 Burlington Road Bedford, MA 01730-1420
	Derek Ruths	druths@ruthsresearch.org	Assistant Professor	Network Dynamics Lab, McGill University	McConnell Engineering Bldg, Rm 318 3480 University Street Montreal, Quebec, Canada H3A 0E9
	Anatoliy Gruzd	gruzd@dal.ca Twitter: @dalprof	Director, Assistant Professor	Social Media Lab, Dahousie University	School of Information Management Kenneth C. Rowe Bldg 6100 University Avenue, Suite 4010 PO BOX 15000 Halifax, NS B3H 4R2 Canada

	Linna Li	linna@geog.ucsb.edu	Postdoctoral Scholar, Researcher	Center for Spatial Studies, University of California- Santa Barbara	Department of Geography 3510B Phelps University of California Santa Barbara, CA 93106-4060, USA
	Kathleen Stewart	kathleen-stewart@uiowa.edu	Associate Professor	University of Iowa	The University of Iowa 305 Jessup Hall Iowa City, IA 52242
	Kristen Summers	ksummers@caci.com	Director, CTO	Advanced Knowledge Solutions Division Group, CACI	Advanced Knowledge Solutions Division Group CACI 4831 Walden Lane Lanham, MD 20706

San Diego Participants (Specialists)

	K. Michael Peddecord	mpeddeco@mail.sdsu.edu	Dr.P.H., Professor Emeritus	San Diego State University, Graduate School of Public Health	6342 Camino Largo, San Diego 92120
	Mark Sawyer	mhsawyer@ucsd.edu	Professor	Professor of Clinical Pediatrics, Division of Infectious Diseases, University of California-San Diego	Department of Pediatrics 3020 Children's Way, MC 5109

Research Team

	Ming-Hsiang Tsou	mtsou@mail.sdsu.edu	Professor	Department of Geography San Diego State University	Department of Geography San Diego State University 5500 Campanile Drive San Diego, CA 92182-4493
	Dipak Gupta	dgupta@mail.sdsu.edu	Professor	Fred J. Jansen Professor of Peace Studies Distinguished Professor in Political Science Program Chair of International Security and Conflict Resolution (ISCOR) San Diego State University	Department of Political Science San Diego State University 5500 Campanile Drive San Diego, CA 92182-4427
	Brian Spitzberg	spitz@mail.sdsu.edu	Professor	Senate Distinguished Professor School of Communication San Diego State University	School of Communication San Diego State University 5500 Campanile Drive San Diego, CA 92182-4560
	Jean Mark Gawron	gawron@mail.sdsu.edu	Professor	Department of Linguistics and Oriental Languages San Diego State University	Department of Linguistics and Oriental Languages San Diego State University 5500 Campanile Drive San Diego, CA 92182-7727
	Li An	lan@mail.sdsu.edu	Associate Professor	Department of Geography San Diego State University	Department of Geography San Diego State University 5500 Campanile Drive San Diego, CA 92182-4493

Student Assistants

Student Name	*Email*	*Level*	*University*	*Department*
Alejandra Coronado	acoronad@rohan.sdsu.edu	Undergraduate	San Diego State University	Geography
Daniel Luscher	lusher@rohan.sdsu.edu	Masters	San Diego State University	Geography
Anna Nagel	annacnagel@gmail.com	Masters	San Diego State University	Public Health
Nicole Stotz	stotz@rohan.sdsu.edu	Masters	San Diego State University	Geography
Sarah Wandersee	wanderse@rohan.sdsu.edu	PhD	San Diego State University	Geography
Ninghua Wang	wangn@rohan.sdsu.edu	PhD	San Diego State University	Geography
Jiue-An Yang	yangj@rohan.sdsu.edu	PhD	San Diego State University	Geography
Tomas Vega	tomasivega@gmail.com	Master	San Diego State University	Political Science

WORKSHOP AGENDA

San Diego State University, San Diego, California
(*Fairfield Inn & Suites San Diego Old Town)*
August 1, 2 (Wednesday, Thursday), 2012

Tuesday, July 31
7:00 pm. Informal gathering for those interested in dinner (coordinated by Dr. Brian Spitzberg. **spitz@mail.sdsu.edu**).

Wednesday, August 1st

*8:00 – Free Breakfast at **Fairfield Inn & Suites San Diego Old Town** (hotel).*
Hot Breakfast served starting at 6:30AM.

9:00 Welcome and Introductions (PI) Ming-Hsiang Tsou

9:10 Background and Introduction to Meeting Goals (Co-PIs) Dipak Gupta, Mark Gawron,
(5 minutes for each Co-PI) Brian Spitzberg, Li An.
Participant self introduction (1 minute per person) Participants

9:40 Overview of CDI project (Year 2) and Research Progress Ming-Hsiang Tsou
(15 mins presentation + 5 mins discussion)

10:00 **GIS Perspective Session (chaired by Tsou):**
From Physical Space to Virtual Space Shih-Lung Shaw
(15 mins presentation + 5 mins Q&A)
Spatiotemporal Event Diffusion: A Formal Model and Framework Kathleen Stewart
(15+5 mins)

10:40 Break 15 minutes [coffee and cookies]

10:55 **Linguistics Perspective Session (chaired by Gawron):**
Diffusion: The problem of tracking ideas (15+5 mins) Mark Gawron
Inferring User Attributes and Extracting Political Sentiment Clayton Fink
from Nigerian Social Media (15+5 mins)
Estimating Community Composition in Twitter and the Real World Derek Ruths
(15+5 mins)
11:55 Group Discussion (10 mins) – open to all participants.

12:10 Lunch - Each Co-PI will bring a group of six people for lunch in old town restaurants (hosted by CDI except alcohol drinks) - Please come back to the hotel by 1:30pm!

1:30 **Homeland Security Perspective Session (chaired by Gupta):**
Tracking Evolution of Narratives: Understanding & Visualizing (20 mins) Dipak Gupta
Social Radar Workflows, Dashboards, and Environments (15+5 mins) Jennifer Mathieu
A Mathematical Approach to Identifying and Forecasting Shifts Les Servi
in the Mood of Social Media Users (15+5 mins)

CyberGeomatic Intelligence – Historical Framework, Michael L. Thomas
Problem Definition and Importance of the Topic (15+5 mins)
2:50 Group Discussion (10 minutes) – open to all participants.

3:00 Break 10 minutes [coffee and cookies]

3:10 **Space-Time Analysis and Integration Session (chaired by An):**

Understanding information landscapes through space-time analysis (15+5 mins) — Li An and Sarah Wandersee
Who tweets and who flickrs? —spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr (15+5 mins) — Linna Li
Vaccine Information and Sentiment Over Space and Time. (15+5 mins) — Anna Nagel, Mark Sawyer, and Michael Peddecord
Group Discussion (10 minutes) – open to all participants

4:20 Break 10 minutes

4:30 **Communication Approaches Session (chaired by Spitzberg):**
Challenges in Integrating MEMES, Networks, and the Noosphere. (15+5 mins) — Brian Spitzberg
Discovery and Visualization of Scholarly Information Diffusion in Twitter Networks (15+5 mins) — Anatoliy Gruzd
Co-Occurrence as an Indicator of Ideas and Sentiment (15+5 mins) — Kristen Summers
Online Deception in AQAP's *Inspire* Magazine: Implications for Crowdsourcing & Gamification (15+5 mins) — Edna Reid
Group Discussion (10 minutes) – open to all participants

6:00 DAY-1 Conclusion and Discussion (Feedbacks from Participants).

6:10 Take group photos.

6:20 Dinner

Thursday, August 2nd

*8:00 – Free Breakfast at **Fairfield Inn & Suites San Diego Old Town** (hotel). Hot Breakfast served starting at 6:30AM.*

9:00 Review of the Day-1 Discussion (4 minutes from each person – 50 minutes)

9:50 Mining and Mine field: a Revolution in Social Science Research — Dipak Gupta

10:10 Break [coffee and snacks]

10:20 CDI Project Challenge #1: **The Veracity of Geolocations in Cyberspace** (IP addresses, GPS, and Tweet self-defined locations). (facilitated by Tsou and An)
10:35 CDI Project Challenge #2: **Sentiment Analysis for 2012 Presidential Election and Collective Identity**. (facilitated by Gawron, Spitzberg, Gupta).

10:50 Charge to the breakout groups (Two Focus Groups – Mission room + Old Town Room).

10:55 – 12:00 Focus Group Discussion:
Locations: Group 1: Mission Room: Geolocation Methods and Spatial Accuracy
Group 2: Old Town Room: Sentiment Analysis Methods (Election)

Each group will be asked to summarize their discussion (15 minutes) and report them in the beginning of the afternoon session.

12:00 Lunch - Each Co-PI will bring a group of six people for lunch in old town restaurants (hosted by CDI except alcohol drinks) - Please come back to the hotel by 1:30pm!

1:30 Reports from the breakout groups. (15 minutes for each group – two groups) + 20 minutes of combined discussion.

2:20 Short Break [coffee and snacks]

2:30 CDI Project Challenge #3: **Space-Time Analysis** and t**he Diffusion of Innovation** (facilitated by An and Tsou)
2:40 CDI Project Challenge #4: **Exploring New Communication Theories and Ontology Models** (facilitated by Spitzberg, Gawron, Gupta).

2:50 Charge to the breakout groups (Two Focus Groups – Mission room + Old Town Room).

3:00 – 3:50 Focus Group Discussion:
Locations: Group 1: Mission Room: The Diffusion of Innovation.
Group 2: Old Town Room: Building Ontologies and communication models

Each group will be asked to summarize their discussion (10 minutes) and report them in the next section.

3*:50 Break [coffee and snacks]*

4:00 Reports from the breakout groups. (10 minutes for each group – two groups) + 10 minutes of combined discussion.

4:30 Research Collaboration Opportunities and the Next Steps.
1. IARPA project (Virginia Tech) (Gupta and Summers) (10 mins)
2. MITRE: Social Radar (Servi and Mathieu) (10 mins)
3. Academic research centers and other organizations (all participants). (10 mins)

5:00 Plenary Discussion: Next Steps. (3 minutes from each participant).

5:40 Project and Workshop evaluation questionnaires (Spitzberg)

6:10 Dinner