Research Node Focus: Duke University/NISS

Duke University and the National Institute of Statistical Sciences (NISS) are working on ways to improve the way U.S. Census and other federal statistical agencies disseminate data. Duke and NISS are focusing on three main areas of research, which include data confidentiality, missing and erroneous data, and combining information across data sources including record linkage.

The Duke University and NISS research node consists of Jerry Reiter, Professor of Statistical Science at Duke University as the PI and Alan Karr, Senior Fellow at NISS and Director of Excellence for Complex Data Analysis at RTI International as the Co-PI. Senior researchers include: Larry Cox, NISS; David Dunson, D. Sunshine Hillygus, V. Joseph Hotz, Fan Li, Seth Sanders and Rebecca Steorts, all from Duke. Postdocs include: Maria De Yoreo, Mauricio Sadinle and former postdocs include: Daniel Manrique-Vallier, now at Indiana University; and Hang Joon Kim, now at University of Cincinnati. Current and former graduate students include: Olanrewaju Akande, Nicole Dalzell, Jingchen Hu, David McClure, Jared Murray, Eve Oh, Tyler Ransom, Tracy Schifeling, Lan Wei, and Jody Heck Wortman.

Federal statistical agencies have an obligation to share data to the public, but they also have an ethical and legal obligation to protect data subjects' privacy and confidentiality. Stripping names, addresses and other direct identifiers typically does not protect confidentiality well enough. As one approach to public use data dissemination, researchers at Duke/NISS are developing novel ways to generate synthetic data, i.e., simulated values, that can protect the identity of businesses or individuals, but still have enough information to be useful for analysis. "We fit these models to try to preserve the overall relationships in the data. Our node is working on developing very flexible models for generating data that capture the key relationships automatically," said Jerry Reiter, PI at Duke University.

In This Issue

- Profile of Duke University and NISS 1
- Shelly Hall Leaving NCRN Coordinating Office 2
- Abowd to Lead Census Bureau Research and Methodology Directorate 3
 - Node News 4
 - NCRN Fall 2015 Meeting 5
 - Publications 6
 - Presentations 6

NCRN Virtual Seminars - 8

As an example, Duke/NISS generated a synthetic version of a subset of data from the Census of Manufactures. Hang Kim, former Duke/NISS postdoc (now at University of Cincinnati) led the team developing this product. "This project illustrates the potential of synthetic data to provide access to data that, currently, one only can access in the research data centers," noted Reiter. Another synthetic data product being developed by the team is an updated version of the Synthetic Longitudinal Business Database (Syn-LBD). The initial SynLBD included synthetic payroll and employment histories for all establishments up to the year 2000. The latest version extends to firm links, NAICS coding, and years up to 2012.



L-R: Jerry Reiter, Maria De Yoreo and Mauricio Sadinle.

The second area that Duke/NISS has focused on is missing and erroneous data, e.g., a daughter is reported to be older than her biological father. The Census Bureau wants to "fix" these data before releasing them to the public. Duke/NISS has developed techniques that have advantages over what agencies typically use. "The method appropriately propagates uncertainty in picking which values are wrong," said Reiter. For example, if a record was reported as a pregnant male who was 42-years-old, the Duke/NISS approach would look at each variable, "pregnant," male" and "42" to help figure out which of these needs to be fixed. "Our methods look at the relationships in the data and try to make changes to see which would be most plausible. So, in this case we would say, 'well, this says the person is 42, so probably being pregnant is reported incorrectly," Reiter explained. The model automatically does this calculation. It is important to emphasize that it is still probably the right

Continued on page 2

Research Node - Duke/NISS (Continued)

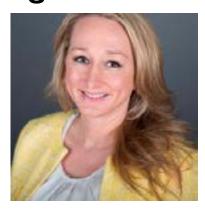
From page 1

answer, but there is always some uncertainty in selecting the variables to change. To represent this uncertainty, the Duke/NISS approach is to create multiple datasets with each one having a plausible correction generated from the model. There is a paper, "Simultaneous Edit-Imputation for Continuous Microdata" in the *Journal of American Statistical Association* (JASA) about this method, including an application to data from the Census of Manufactures. There is a companion article that is currently under review that considers categorical data.

The third area of focus is on data integration. "This is a major focus during the remaining period of the grant," said Reiter. In this area of research, Duke and NISS are working on data integration and record linkage. Record linkage arises when a researcher is looking at two or more datasets and deciding which records match together or not. The most challenging cases arise when there is no definitive identifier to inform that decision, such as a Social Security number. For example, there may be a "Jim Smith" and a "James Smith" in the two databases. One has to decide if these are two distinct people, or if they are the same person. Just like the missing/erroneous data, the Duke/NISS researchers are creating multiple datasets to propagate uncertainty to the users. For example, if both records say this person lives at the same address, the uncertainty is lower than if there were two different addresses. Reiter said, "The main thing we have been focusing on is, how do you account for uncertainty in your ultimate inference when you have that inexact linking?" So, in our original example of "Jim Smith" and "James Smith," if one dataset has the person's education status and the other dataset has income level and we want to link the two together, but we are still not 100% sure that this is the same person, we want to incorporate the uncertainty into these matches. "So much like we do for the missing or synthetic data, we wind up creating multiple datasets and build a statistical model that more or less puts probability on whether there is a link between two records, and then create plausible matches by looking across everybody's links and figuring out who gets matched with whom," explained Reiter.

Shelly Hall Leaving NCRN and the NCRN Coordinating Office

Shelly Hall, project assistant to Lars Vilhuber for the NCRN Coordinating Office at Cornell University, has accepted a new position as pre-award specialist in Cornell University's College of Engineering. Shelly's last day with NCRN Coordinating Office was Friday, December 18th, 2015. Shelly had worked with the ILR School, the Economics Department and



Shelly Hall is leaving the NCRN Coordinating Office.

the Labor Dynamics Institute over the past seven years, providing the department and faculty with organizational and administrative skills that will be missed by all. We wish her well in her new position. Until a replacement for Shelly is hired, Sophia Harmon will assist Lars Vilhuber in handling many aspects of the Coordinating Office's tasks.

The Coordinating Office continues to be reachable at info@ncrn.info for any questions, suggestions, and comments.

Renee Ellis Promoted, New Person to be NCRN Liaison

U.S. Census Bureau NCRN Liaison Renee Ellis announced that she has accepted a promotion at the U.S. Census. Therefore she will be stepping down from her position as the NCRN Liaison for the Census Bureau. A new liaison will be announced soon.



Renee Ellis.

Abowd to Lead Census Bureau Research and Methodology Directorate

Cornell University professor John M. Abowd will join the U.S. Census Bureau through an interagency personnel agreement as the new associate director for research and methodology and chief scientist, effective June 1, 2016. The announcement was made by Census Bureau Director John H. Thompson on December 10, 2015.

The associate director for research and methodology leads a directorate of research centers, each devoted to domains of investigation important to the future of social and economic statistics. In that regard, the position leads key capacities for innovation across all statistical programs of the Census Bureau. The Research and Methodology Directorate includes the Center for Adaptive Design, Center for Administrative Records Research and Applications, Center for Disclosure Avoidance Research, Center for Economic Studies, Center for Statistical Research and Methodology, Center for Survey Measurement and the new Center for Big Data Research and Applications.

"I'm thrilled to have Dr. Abowd coming on board to help lead critical work to modernize the Census Bureau's operations and products," Thompson said. "John is an internationally renowned scholar and a great addition to lead our research and methodology efforts. His research on privacy and statistical disclosure limitation are particularly relevant as we work to improve the utility of our data products to our users and honor our commitment to maintain the confidentiality of the data the American people and businesses provide us."

Abowd joined the faculty of Cornell University's School of Industrial and Labor Relations in 1987. He is currently the Edmund Ezra Day Professor of economics, statistics and information science. He began his long association with the Census Bureau in 1998 when he joined the team of distinguished senior research fellows that helped found the Longitudinal Employer-Household Dynamics program. He has served continuously since 1998 as a scientific adviser to that program and others at the bureau. Since 2011, he has been the principal investigator for the Cornell University node of the National Science Foundation-Census Research Network (NCRN) and co-principal investigator of its coordinating office.

Abowd is a fellow and past president of the Society of Labor Economists. He is also a fellow of the American Statistical Association and the Econometric Society, as well as



John M. Abowd is joining the U.S. Census Bureau.

an elected member of the International Statistical Institute. He currently serves on the National Research Council's Committee on National Statistics and the American Economic Association's Committee on Economic Statistics.

"This is a period of enormous methodological challenges for the Census Bureau and all national statistical agencies that will require innovative redesign and reengineering of the bureau's flagship activities," Abowd said. "I am excited to join Director John Thompson, Deputy Director Nancy Potok and the associate directors in marshaling the bureau's talented scientific professionals to meet these challenges."

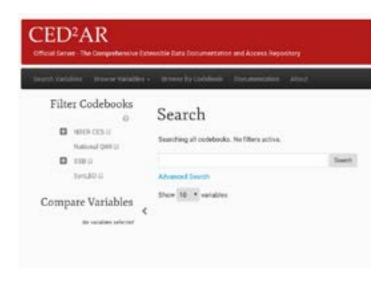
Abowd will be withdrawing as Principal Investigator from the NCRN grants at Cornell (replaced by co-PI Lars Vilhuber) and at the Coordinating Office.

Node News

Crowdsourcing metadata editor now in beta release at NCRN Cornell

User feedback has been a critical component of the Cornell Synthetic Data Server since its start, and the incorporation of user feedback into the synthetic data sets hosted on the Synthetic Data Server has contributed to their improvement.

Cornell released new tools to assist users of the Synthetic Data Server in the improvement of the datasets they used. In particular, the team made tools for improving documentation. Documentation for the SIPP Synthetic Beta and the Synthetic LBD have been available at https://www2.ncrn.cornell.edu/ced2ar-web/search since 2014. These documents were developed entirely by data curators. Users can now access the wiki version by selecting a codebook, for instance the SSB v6.02 codebook, and clicking "View crowdsourced contributions" on any page.



Cornell Releases JAGS 4.0 RPM Packages

Cornell is now providing packaging support for JAGS (Just Another Gibbs Sampler). Version 4.0 was recently released, and binary packages are now available. More details at https://www.ncrn.cornell.edu/software/jags-for-rhel-centos-opensuse/.

Steorts Presents at EmTech MIT

Rebecca Steorts, Assistant Professor at Duke University, recently spoke at the "Meet the Innovators Under 35" talk at EmTech MIT. EmTech is a conference that *MIT Technology Review* holds to help discover future trends and understand the technologies that will drive the new global economy. It is where technology, business, and culture converge. EmTech was held in Cambridge November 2-4. Rebecca's talk was "Big Data Could Cut Through the Fog of War."

Her speech highlighted how she uses record linkage, which uses multiple noisy databases to remove duplicate entries. The methodology she is using allows her to run 300,000 records in 10 minutes and the accuracy is 99 percent. Other methods' accuracy is between 50-70 percent and can take days or even weeks to complete.

You can watch her presentation here.



Rebecca Steorts.

NCRN Fall 2015 Meeting

The NCRN Fall 2015 Meeting took place at the Bureau of Labor Statistics Conference Center in Washington DC, and at the Census Bureau's headquarters in Suitland, MD, on December 14 and 15, 2015.

On December 14, about 90 participants from many different federal agencies (Bureau of Labor Statistics, Census Bureau, Centers for Disease Control and Prevention, Social Security Administration, Small Business Administration, Internal Revenue Service, the National Institutes of Health and the National Science Foundation), all NCRN nodes, as well as a number of private and academic research institutions, attended the meeting.

Lars Vilhuber, Principal Investigator of the NCRN Coordinating Office, and Erica Groshen, the Commissioner of the Bureau of Labor Statistics, made opening remarks. The two sessions that followed, on "Confidentiality" before lunch, and on "Small domain estimation and visualization of uncertainty in small area data", dwelled on topics that are of interest to a broad range of statistical applications in the federal statistical system and in the private sector. John Abowd (Cornell NCRN) presented a new method for "Formal privacy protection for data products combining individual and employer frames," the result of joint work with



Participants listening to a session at the NCRN Fall Meeting.

Census Bureau staff and academics from Cornell and Duke University. Nicolas Kim, graduate student at Carnegie Mellon University, presented his research on "The effect of data swapping on contingency table analysis". Jerry Reiter (Duke NCRN) addressed the issue of "Simultaneous editimputation and synthetic data generation for establishment microdata," joint work with former NCRN graduate student Hang Kim and his fellow PI, Alan Karr, now at RTI. The method aims to produce better public-use data from the Census of Manufactures. John Eltinge, Associate Commissioner for Office of Survey Methods Research, summarized and commented on the presentations from the perspective of the possible implementation in a statistical agency. He noted the broader context for agencies, which need to balance quality, cost, risk, and utility to multiple stakeholders, and in particular in the context of statistics as "public goods". His comments, and all the presentations, can be viewed on the website for the NCRN Fall 2015 Meetings.

The afternoon sessions focused on scalable small-domain estimation, as well as the challenge on how to visualize

uncertainty. Noel Cressie (Missouri NCRN and University of Wollongong) reported on work with Australian "Census Data on Mean Usual Weekly Income: Regression and Simultaneous Autoregression when the Dataset is Large". Scott Holan (Missouri NCRN) considered the issue of "Spatio-temporal change of support with application to American Community Survey multi-year period estimates", although the issue can be applied to many other contexts where geographic and temporal granularity of statistics varies in a hierarchical fashion. Jason Jurjevich (Portland State University and member of the Colorado/Tennessee NCRN node) presented the latest results from their node's mapping experiments, in "Navigating ACS Data Uncertainty: Insights from Mapping Experiments with Urban Planners". The project is of particular relevance to multiple agencies. who struggle with the challenge of presenting users with the uncertainty inherent in statistics (here, urban planners, but any kind of public user faces the same problems). Others, including Northwestern NCRN PI Charles Manski, have argued that statistical agencies need to mitigate misinterpretation of published statistics by communicating uncertainty to the public, and Jason and his collaborators have worked to address that issue. Carson Farmer (University of Colorado at Boulder) concluded the presentations by "Exploring a scalable framework for spatially-local regression analysis". Connie Citro (Committee on National Statistics) discussed and commented the papers from the afternoon session. She highlighted in particular the need and challenges of migrating from theory or demonstration to the production world of statistical agencies. Her comments, and all the presentations, can be viewed on the website for the NCRN Fall 2015 Meetings.

On December 15, NCRN participants and Census Bureau staff met in a number of parallel breakout sessions to discuss topics relevant to the upcoming 2020 decennial census. Topics included internet self-response, re-engineering of address canvassing, geographic and demographic targeting and market segmentation, the use of administrative records for edit and imputation of data, the re-engineering of field operations, cost-quality trade-off analysis, internet survey geolocation studies. The NCRN Principal Investigators also met with Census Bureau Director Thompson and staff from the Research and Methodology directorate to discuss future directions and possibilities of interaction between academic researchers and the Census Bureau.



Noel Cressie, Co-PI of the Missouri node, speaking to attendees at the NCRN Fall 2015 Meeting.

Publications

The following are the most recent additions to publications produced by the research nodes within NCRN. A comprehensive list can be found here. (http://www.ncrn.info/documents/bibliographies)

Abowd, John M., Kevin L. McKinney, and Ian M. Schmutte. *Modeling Endogenous Mobility in Wage Determination*. Cornell University Preprint 1813:40306, 2015, available at http://hdl.handle.net/1813/40306.

Abowd, John, and Ian Schmutte. *Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods*. Cornell University Preprint 1813:39081, 2015, available at http://hdl.handle.net/1813/39081.

Eck, A., S. Leen-Kiat, A. L. McCutcheon, J.D. Smyth, and R.F. Belli. "Understanding the Human Condition through Survey Informatics." *IEEE Computer* 48, no. 11 (2015): 112-116. DOI: 10.1109/MC.2015.327.

Fosdick, B. K., M. De Yoreo, and J. P. Reiter. "Categorical data fusion using auxiliary information." *ArXiv* (2015), available at http://arxiv.org/pdf/1506.05886.pdf.

H. Shaefer, Luke. *Introduction to The Survey of Income and Program Participation (SIPP)*. University of Michigan Preprint 1813:40169, 2015, available at http://hdl.handle.net/1813/40169.

Shrivastava, Anshumali. *Probabilistic Hashing Techniques For Big Data*, *Computer Science*. Vol. Ph.D. Cornell University Dissertation, 2015, available at https://ecommons.cornell.edu/handle/1813/40886.

Akande, O., Fan Li, and J. P. Reiter. "An empirical comparison of multiple imputation methods for categorical data." arXiv (2015).

De Yoreo, M., J. P. Reiter, and D. S. Hillygus. "Nonparametric Bayesian models with focused clustering for mixed ordinal and nominal data." arXiv (2015).

T. Paiva and J. P. Reiter, "Stop or continue data collection: A nonignorable missing data approach for continuous variables," arXiv (2015)

Paiva, T., and J.P. Reiter. "Stop or continue data collection: A nonignorable missing data approach for continuous variables." ArXiv (2015)

McClure, D.. Relaxations of differential privacy and risk utility evaluations of synthetic data and fidelity measures, *Duke University Statistics Department*. Vol. PhD. dissertation, 2015.

Steorts, R. C. "Entity resolution with empirically motivated priors." Bayesian Analysis 10(5) 2015

Presentations

On September 19-22, the International Conference on Total Survey Error was held in Baltimore. **Kristen Olson**, **Jolene Smyth** and **Antje Kirchner** from University of Nebraska-Lincoln presented "Do Interviewers with High Cooperation Rates Behave Differently? Interviewer Cooperation Rates and Interview Behaviors." **Jerry Reiter**, Duke University, presented "Relationships between Data Quality and Confidentiality" and **Alan Karr**, RTI International, gave a talk on "The Role of Statistical Disclosure Limitation in Total Survey Error" at the conference.

Scott Holan, Missouri, presented "Multivariate Spatio-Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics," at the NBER-NSF Time Series Conference, Vienna, Austria on September 25.

Christopher Wilke, Missouri, presented "Efficient parameterizations and optimal support for multiscale multivariate spatio-temporal data" as an invited speaker at the Colorado School of Mines in Golden, Colorado on October 2.

Maria De Yoreo, Duke University, presented "Incorporating Conditionally Representative Auxiliary Information in Data Fusion" on October 7 in the NCRN Virtual Seminar (a recording is available at http://is.gd/220ABM)

Presentations (Continued)

On October 16, 2015, the Geospatial Statistics, Tools, Data, Practices, Opportunities and Challenges in the Federal Agencies Workshop took place. Featured talks and speakers included: "Recent Developments for Space-Time Modeling of Public-Use Federal Data" by Jonathan Bradley and "Recent Advances and Future Directions in Spatial and Spatio-Temporal Statistics for Federal Data: Academic Partnerships and the NCRN" by Scott Holan, both from the University of Missouri research node.

Jerry Reiter, Duke, presented, "Bayesian modeling and multiple imputation to simultaneously handle missing and erroneous values and protect confidentiality in Census Bureau data," at the 65th Birthday Conference for Rod Little, University of Michigan, October 30, 2015.

November 4, 2015, **Robert Colosi** from the U.S Census Bureau presented the "2020 Census Operational Plan" at the NCRN Virtual Seminar (a recording is available at http://is.gd/EyeiwB).

November 19, 2015, "Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models with Local Dependence", **Jared Murray**, Carnegie Mellon University.

November 20-21, 2015, **Beth Cochran, Kristen Olson** and **Jolene D. Smyth**, University of Nebraska, presented "When to Repeat: Interviewer and Respondent Behavior on Battery Questions," at the meeting of the Midwest Association for Public Opinion Research in Chicago, Illinois. **Antje Kirchner, Kristen Olson** and **Jolene D. Smyth** also presented "Does laughter mean the respondent is friendly? Examining interviewer post-survey evaluations and respondent behaviors" at that same conference.

The Federal Committee on Statistical Methodology (FCSM) 2015 Research Conference took place December 1-3 in Washington DC. Members of several NCRN nodes gave talks.

- Concurrent Session B-3: "Total Variability Measures for Selected Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in OnTheMap," Kevin McKinney (U.S. Census Bureau) Andrew Green (Cornell University and U.S. Census Bureau) Lars Vilhuber (Cornell University and U.S. Census Bureau)
 John Abowd (Cornell University and U.S. Census Bureau)
- Concurrent Session C-1: "Two Perspectives on Commuting and Workplace: A Microdata Comparison of Home to Work Flows Across Linked Survey and Administrative Files" **Andrew Green** (U.S. Census Bureau, Cornell University) Mark Kutzbach (U.S. Census Bureau) **Lars Vilhuber** (U.S. Census Bureau, Cornell University)
- Concurrent Session D-2: "Crowdsourcing Codebook Enhancements: A DDI-based Approach" Benjamin
 Perry (Cornell University) Venkata Kambhampaty (Cornell University) Kyle Brumsted (McGill University) Lars Vilhuber (Cornell University) William Block (Cornell University)
- Concurrent Session E-1: "Within-Industry Productivity Dispersion and Imputation for Missing Data" T. Kirk White (U.S. Census Bureau) Jerome P. Reiter (Duke University) Amil Petrin (University of Minnesota and NBER)
- Concurrent Session I-3: "Formal Privacy Protection for Data Products Combining Individual and Employer Frames Ashwin Machanavajjhala (Duke University) Samuel Haney (Duke University) Matthew Graham (U.S. Census Bureau) Mark Kutzbach (U.S. Census Bureau) John Abowd (U.S. Census Bureau, Cornell University)

Jerry Reiter presented, "Data access and data confidentiality: Where are we and where are we going?" at the Commissioner's Invited Lecture Series, Bureau of Labor Statistics, December 15 2015.

December 19-20, 2015 **John Abowd** presented "Labor Market Measurement in the 21st Century" at the 2015 International Symposium on Contemporary Labor Economics sponsored by the Wang Yanan Institute for Studies in Economics (WISE) at Xiamen University

Recent NCRN Virtual Seminar

January 6, 2016. "Synthetic establishment and firm data"

"Synthetic Data Generation for Firm Links" (Saki Kinney, NISS)

In most countries, national statistical agencies do not release establishment-level business microdata, because doing so represents too large a risk to establishments' confidentiality. Agencies potentially can manage these risks by releasing synthetic microdata, i.e., individual establishment records simulated from statistical models designed to mimic the joint distribution of the underlying observed data. Previously, we used this approach to generate a public-use version---now available for public use---of the U.S. Census Bureau's Longitudinal Business Database (LBD), a longitudinal census of establishments dating back to 1976. While the synthetic LBD has proven to be a useful product, we now seek to improve and expand it by using new synthesis models and adding features. This paper describes our efforts to create the second generation of the SynLBD, including synthesis procedures that we believe could be replicated in other contexts.

"Assessing the Data Quality of Public Use Tabulations Produced from Synthetic Data: Synthetic Business Dynamics Statistics" (Lars Vilhuber, Cornell)

We describe and analyze a method that blends records from both observed and synthetic microdata into public-use tabulations on establishment statistics. The resulting tables use synthetic data only in potentially sensitive cells. We describe different algorithms, and present preliminary results when applied to the Census Bureau's Business Dynamics Statistics and Synthetic Longitudinal Business Database, highlighting accuracy and protection afforded by the method when compared to existing public-use tabulations (with suppressions).

Upcoming NCRN Virtual Seminars

Here are the upcoming virtual seminars for NCRN. All talks take place from 3-4 p.m. Eastern Standard Time. Updated information is available at:

http://www.ncrn.info/events/virtual-seminar

February 3, 2016 Microclustering: When the Cluster Sizes Grow Sublinearly with the Data Set. (Rebecca Steorts, Duke University)

Most generative models for clustering implicitly assume that the number of data points in each cluster grows linearly with the total number of data points. Finite mixture models, Dirichlet process mixture models, and Pitman--Yor process mixture models make this assumption, as do all other infinitely exchangeable clustering models. However, for some tasks, this assumption is undesirable. For example, when performing entity resolution, the size of each cluster is often unrelated to the size of the data set. Consequently, each cluster contains a negligible fraction of the total number of data points. Such tasks therefore require models that yield clusters whose sizes grow sublinearly with the size of the data set. We address this requirement by defining the microclustering property and introducing a new model that exhibits this property. We compare this model to several commonly used clustering models by checking model fit using real and simulated data sets.

March 2, 2016 "The effect of question and questionnaire characteristics on interviewer and respondent behaviors in CATI surveys."

Kirsten Olson (Nebraska)

The NCRN Newsletter is published quarterly by the NCRN Coordinating Office. You may reach us at: 275 Ives Hall, Cornell University, Ithaca, NY 14853 (607) 255-2744 info@ncrn.info

www.ncrn.info

Jamie Nunnelly - Managing Editor