NDHS Program Overview

Anjela Govan
North Carolina State University
SAMSI Agricultural Group

NDHS Working Groups

- Anomaly Detection
- Agricultural Systems
- Social Networks
- Data Confidentiality

Anomaly Detection Group

- 35 members – estimated average weekly attendance 20, including 4 by phone
- 24 weekly meetings – research presentations with discussion
  - 19 different presenters
- Midyear workshop at National Center for Health Statistics in Washington DC
  - 3 keynote speakers, 5 other invited speakers
- Several research projects coming from these activities will continue into the future

Anomaly Detection Group

- Main topics of study
  - Scan statistics – detecting anomalous clustering (of crimes, email messages, etc.)
  - Change point analysis - detecting temporal changes in patterns of behavior
  - Syndromic surveillance – ways to detect disease outbreaks
  - Regression diagnostics and outlier detection – which suspicious features are true anomalies?
- Working papers being prepared (see website)
  - Using mixtures to detect spurious observations (F. Vera, J. Lynch, D. Dickey)
  - Bayesian multiple testing (G. Datta, D. Banks)
  - Sequential tests for unit roots in the first-order autoregressive model (Y. Mei, D. Dickey)

Social Networks Group

- Networks arise in many fields; one wants models that describe which connections are likely to get formed or broken.

- This group looked at, among others, Enron email networks, elephant social networks, and networks based on intelligent agent models.

Social Networks Group

- The main challenge is to develop realistic models for network dynamics. Previous theory has focused only upon static networks.
- Some main issues involve the use of latent variables, clique structures as building blocks, feedback loops, and the development of principled methods for assessing the fit of network models.

Social Networks Group

The main accomplishments were:

- An National Academies workshop
- An upcoming edited volume
- Various papers and presentations.

Data Confidentiality Group
What's the Problem?

- People are very concerned about their "personal information." Some examples:
  - NSA telephone call databases
  - Identity theft
  - Could health records be used to deny employment?
- Government agencies collect a lot of data
- There are legitimate uses of the data
  - Counter-terrorism
  - Policy formulation and evaluation
- How can (or *can*) these competing interests be satisfied?

Data Confidentiality Group Statistical Disclosure Limitation

- Goal: Make released data or information (data summaries) both safe and useful.
- Huge range of strategies
  - Restrict access to data
  - Coarsen the data: tell the truth, but not the whole truth. Example: change date of birth to age in years.
  - Alter the data: tell a “near truth.” Example: swap attribute values.
  - And many more

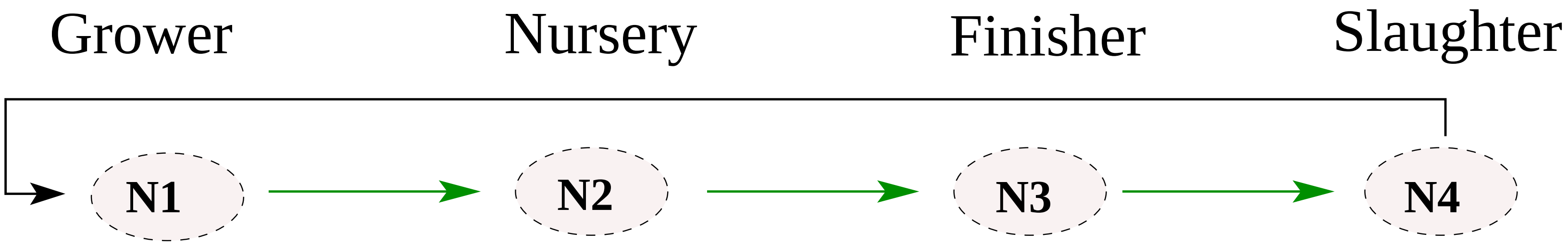
What Did the Working Group Work On?

- Ways to quantify safety and usefulness of released information
- New strategies for statistical disclosure limitation
- Combining statistical disclosure limitation strategies to obtain something that is better than either alone
- Algorithms and software to do statistical analyses of distributed databases

Agricultural Systems Group

- Model the swine production network.
- Model spread of Foot and Mouth Disease through the production network in North Carolina.

Swine production network



Stochastic Model

$$\lambda_1(\mathbf{X}) = q(X_1 - 1, X_2 + 1, X_3, X_4) = k_1 X_1 (L_2 - X_2)_+$$

$$\lambda_2(\mathbf{X}) = q(X_1, X_2 - 1, X_3 + 1, X_4) = k_2 X_2 (L_3 - X_3)_+$$

$$\lambda_3(\mathbf{X}) = q(X_1, X_2, X_3 - 1, X_4 + 1) = k_3 X_3 (L_4 - X_4)_+$$

$$\lambda_4(\mathbf{X}) = q(X_1 + 1, X_2, X_3, X_4 - 1) = k_4 \min(X_4, L_1)$$

Deterministic Model

$$dy_1/dt = -k_1 y_1 (L_2 - y_2)_+ + k_4 \min(y_4, L_1)$$

$$dy_2/dt = -k_2 y_2 (L_3 - y_3)_+ + k_1 y_1 (L_2 - y_2)_+$$

$$dy_3/dt = -k_3 y_3 (L_4 - y_4)_+ + k_2 y_2 (L_3 - y_3)_+$$

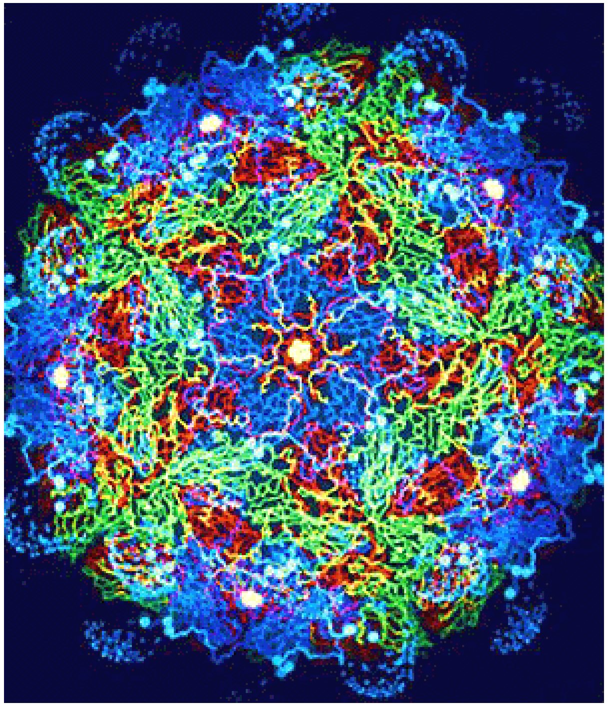
$$dy_4/dt = -k_4 \min(y_4, L_1) + k_3 y_3 (L_4 - y_4)_+$$

Sensitivity Analysis

- Determining right parameter values is hard.
- Which parameters effect the system the most?
- Determine the sensitivity of the model to its various parameters.

Foot and Mouth Disease

- Foot and Mouth is a highly contagious viral disease of cloven-hoofed animals.
- Foot and Mouth has a devastating effect on the export and domestic market of the infected country.
- North America is currently Foot and Mouth free.

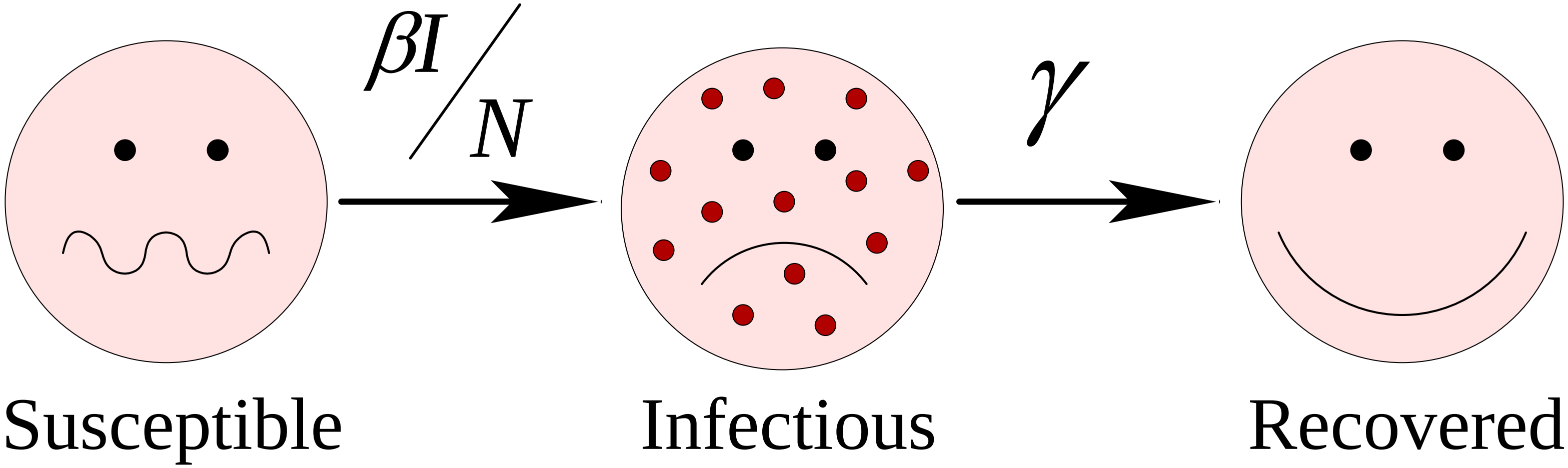


Basic Disease Model SIR

- **Susceptible** (S) – **Infectious** (I) – **Recovered** (R)

- Start with a well-mixed population of size N
- Assume permanent immunity after recovery
- No death, no birth

SIR Equations



$$dS/dt = -\beta SI/N$$

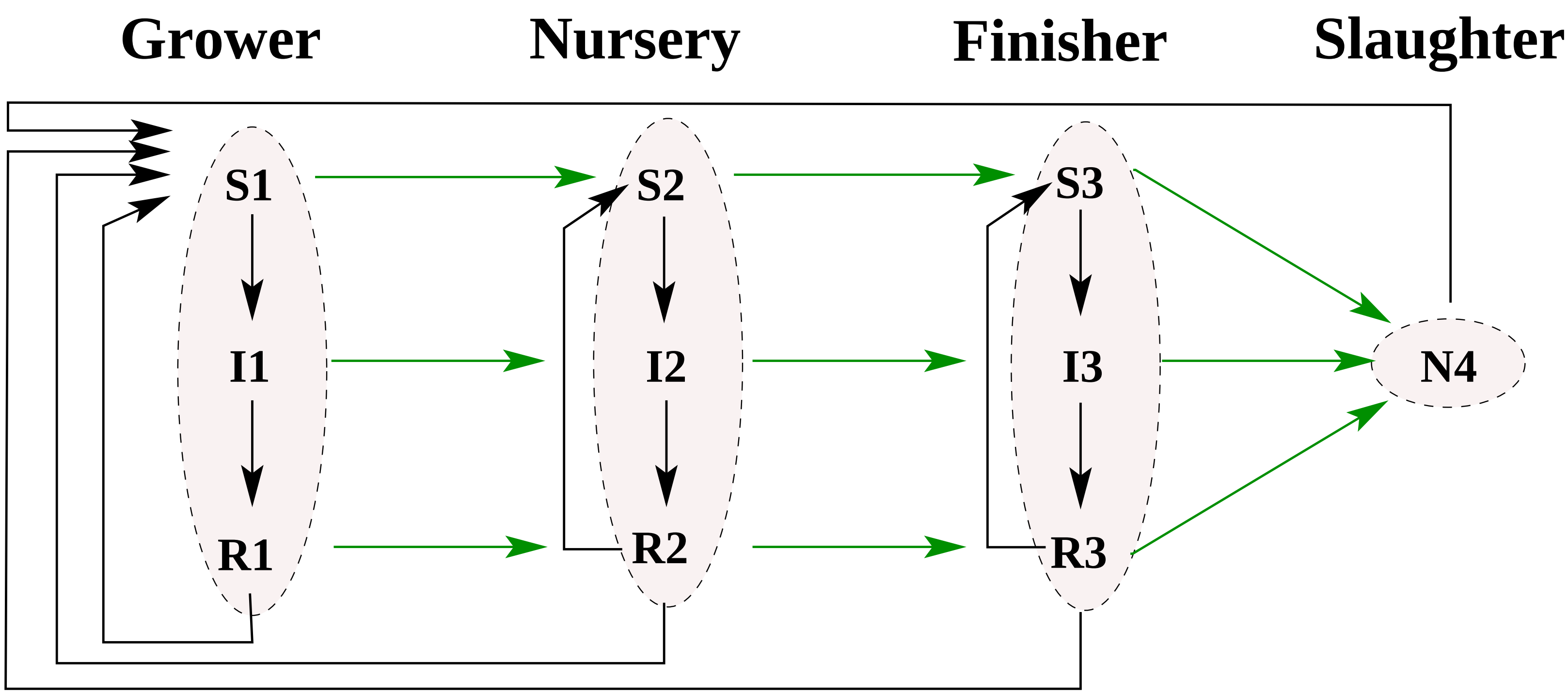
$$dI/dt = \beta SI/N - \gamma I$$

$$dR/dt = \gamma I$$

Assumptions for SIR in Ag network

- Pigs are born healthy.
- There is no infection or recovery during transport between the nodes.
- There is no infection in the slaughter node.
- No human intervention.
- Infected pigs recover or die.
- Recovered pigs have temporary immunity.

Agricultural model with SIR



Ag Model with SIR

- Node 1 (Sows)

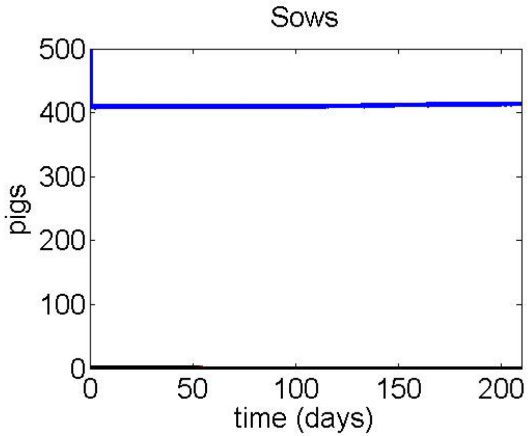
$$dS_1/dt = -\beta_1 S_1 I_1/(N_1) - k_1 S_1 (L_2 - N_2)_+ +$$

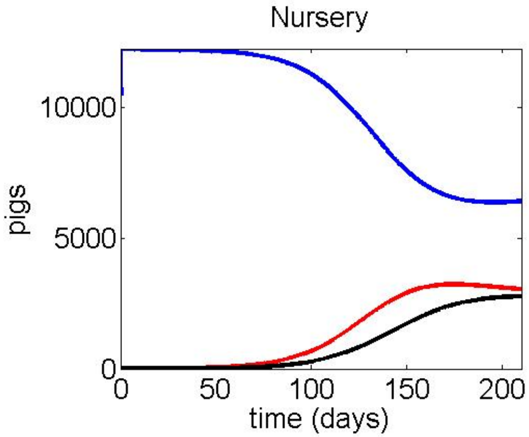
$$+ k_4 \min(y_4, L_1) + \delta_1 R_1 + \delta_2 R_2 + \delta_3 R_3 + \rho_1 R_1;$$

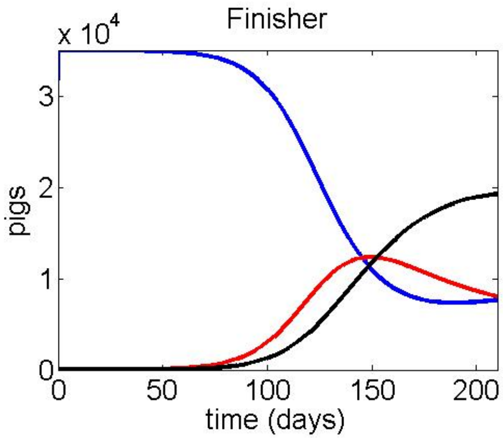
$$dI_1/dt = (\beta_1 S_1 I_1)/(N_1) - \gamma_1 I_1 - k_1 I_1 (L_2 - N_2)_+;$$

$$dR_1/dt = \gamma_1 I_1 - k_1 R_1 (L_2 - N_2)_+ - \delta_1 R_1 - \rho_1 R_1;$$

Simulation


Sows
500
400
300
200
100
0
pigs
0
50
100
150
200
time (days)


Nursery
10000
5000
0
pigs
0
50
100
150
200
time (days)


Finisher
x 10^4
3
2
1
0
pigs
0
50
100
150
200
time (days)


Slaughter
250
200
150
100
50
0
pigs
0
50
100
150
200
time (days)

Challenges

- Disease Parameters
  - Infectious rate β is hard to compute.
  - Recovery rate γ and loss of immunity rate ρ are computed for farms.
  - No age distinction for γ and ρ.

Conclusion

- Use sensitivity analysis on Ag network with SIR.
- Incorporate human intervention.
  - slaughter
  - vaccination

Acknowledgments
Social Networks Group

- Deepak Agarwal (Yahoo!);
- Edo Airoldi (CMU);
- David Banks (Duke),
- Hugh Chipman (Acadia);
- Myron Katzoff (NCHS);
- Mike Last (NISS);
- Eric Vance (Duke);
- Rob Warren (Waterloo).

Acknowledgments
Social Networks Group

- Faculty types: Larry Cox (National Center for Health Statistics=NCHS), Joe Fred Gonzalez (NCHS), Alan Karr (NISS), Jim Lynch (South Carolina), Bahjat Qaqish (UNC), Jerry Reiter (Duke)
- Postdocs: Lisa Denogean (SAMSI), Anna Oganian (NISS), Francisco Vera (SAMSI/NISS), Mi-Ja Woo (NISS)
- Students: Joyee Ghosh (Duke), Robin Mitra (Duke), Sati Sarker (Duke)

Acknowledgments
Agricultural Systems Group

- Ping Bai, Statistics, UNC
- Thomas Banks, Mathematics, NCSU
- Sava Dediu, SAMSI/NCSU
- Anjela Govan, Mathematics, NCSU
- Michael Last, NISS
- Alun Lloyd, Biomathematics, NCSU
- Hoan Nguyen, Mathematics, NCSU
- Mette Olufsen, Mathematics, NCSU
- Grzegorz Rempala, University of Louisville
- Barrett Slenning, NCSU Vet School

Thank you.