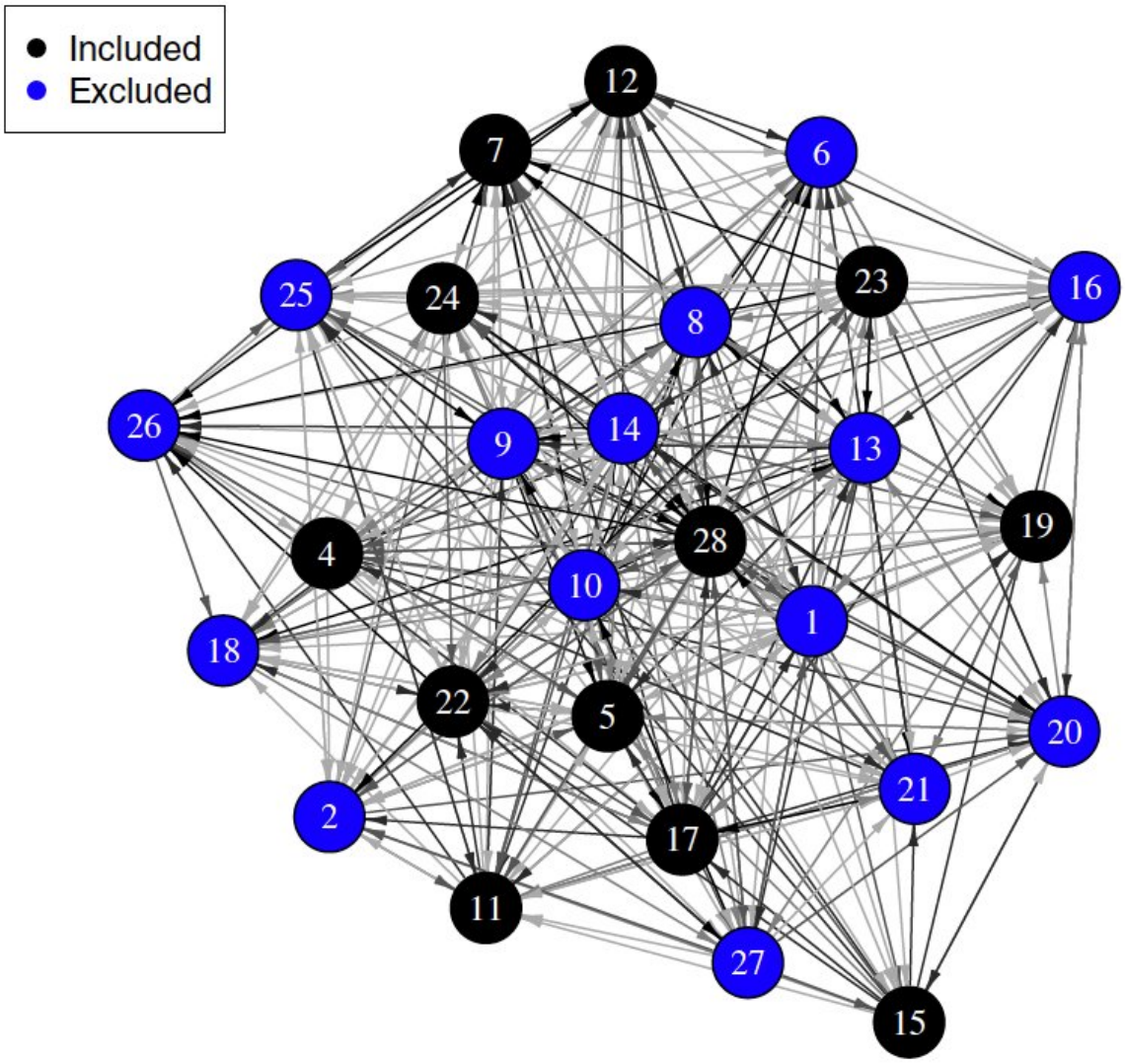
An unbiased approach for sifting through big data

February 2 2018



Nodes and lines represent the health-related variables and the strength of interdependence between two variables respectively. MENet helps build the

Optimal Information Network (OIN) which indicates the most useful information to accurately characterize systemic health. Credit: Servadio J.L. and Convertino M, *Science Advances*, Feb. 2, 2018.

Researchers have developed a complex system model to evaluate the health of populations in some U.S. cities based only on the most significant variables expressed in available data. Their unbiased, network-based probabilistic approach to mine big data could be used to assess other complex systems, such as ranking universities or evaluating ocean sustainability.

Sifting through large amounts of data to determine which variables to use for the assessment of things like the health of a city's population is challenging. Researchers often choose these variables based on their personal experience. They might decide that adult obesity rates, mortality rates, and life expectancy are important variables for calculating a generalized metric of the residents' overall health. But are these the best variables to use? Are there other more important ones to consider?

Matteo Convertino of Hokkaido University in Japan and Joseph Servadio of the University of Minnesota in the U.S. have introduced a novel probabilistic method that allows the visualization of the relationships between variables in big data for complex systems. The approach is based on "maximum transfer entropy," which probabilistically measures the strength of relationships between multiple variables over time.

Using this method, Convertino and Servadio mined a large quantity of health data in the U.S. to build a "maximum entropy network" (MENet): a model composed of nodes representing health-related variables, and

lines connecting the variables. The lines are darker the stronger the interdependence between two variables. This allowed the researchers to build an optimal information network (OIN) by choosing the variables that had the most practical relevance for assessing the health status of populations in 26 U.S. cities from 2011 to 2014. By combining the data from each selected variable, the researchers were able to compute an "integrated health value" for each city. The higher the number, the less healthy a city's population.

They found that some cities, such as Detroit, had high values indicating poor overall health during that timeframe. Others, such as San Francisco, had low values, indicating more favorable health outcomes. Some cities, such as Philadelphia, showed high variability over the four-year period. Cross-sectional comparisons showed tendencies for California cities to score better than other parts of the country. Also, Midwestern cities, including Denver, Minneapolis and Chicago, appeared to perform poorly compared to other regions, contrary to national city rankings.

Convertino believes that methods like this, fed by large datasets and analysed via automated stochastic computer models, could be used to optimize research and practice; for example, for guiding optimal decisions about health. "These tools can be used by any country, at any administrative level, to process data in real-time and help personalize medical efforts," says Convertino.

But it is not just for health data. "The model can be applied to any complex system to determine their optimal information network, in fields from ecology and biology to finance and technology. Untangling their complexities and developing unbiased systemic indicators can help improve decision-making processes," Convertino added.

More information: J.L. Servadio at University of Minnesota School of Public Health in Minneapolis, MN el al., "Optimal information

networks: Application for data-driven integrated health in populations," *Science Advances* (2018). advances.sciencemag.org/content/4/2/e1701088

Provided by Hokkaido University