Submission for The Design Science Award of the INFORMS Information Systems Society.

December 4, 2009

Project Title: Social Network-based Marketing Systems

Nominee contact:

Professor Foster Provost
Information Systems Group, IOMS Dept.
Stern School of Business, NYU
fprovost@stern.nyu.edu
212-998-0806

Nominee team

NYU/Stern (current or former): Brian Dalessandro (now at Media6degrees), Shawndra Hill (now at UPenn/Wharton), Sofus Macskassy (now at Fetch Technologies), Foster Provost, Xiaohan Zhang

AT&T Research: Chris Volinsky

Coriolis Ventures: Rod Hook, Alan Murray

Special thanks to Media6degrees, to the Marketing Science Institute for a research award, to the New York Software Industry Association for an award, to the U.S. Government for substantial funding, and to our many colleagues who provided input over many years, especially to Ted Senator for stimulating serious design science work on modeling with network data during his tenure at DARPA.

The goal of this long-term project was to develop techniques and systems for taking advantage of massive data on social networks to improve on-line advertising, direct marketing, and other predictive modeling tasks. This project developed a suite of techniques for predictive modeling with social network data. These techniques are described in a series of published papers, with different but closely related focuses (please see the accompanying papers for details). The various techniques have been rigorously and extensively evaluated on benchmark data sets for relational classification (not marketing data), a real application of direct marketing, and a real application of on-line brand advertising. This nomination includes four papers that track the development of this line of design science work.

The main idea underlying this line of work is that information systems to support predictive modeling, and in particular marketing/advertising, can take advantage of relational autocorrelation among neighbors in a social network with respect to characteristics such as brand affinity and propensity to purchase. This is possible if the systems can access data on the social network, or on a reasonable proxy to the social network. The long-term project first developed basic technologies and ideas outside the context of marketing applications, and then moved to a targeted marketing application, and then to a large-scale on-line advertising application. Each of these steps in the progression made a contribution to the research literature: all but the first paper are published in top venues; each of the three papers published prior to 2009 has received more than 50 Google Scholar citations. The work described in the 2006 paper in *Statistical Science* has received a large amount of interest in marketing-related blogs, and was called an "influential paper" in an article on social-network marketing in the New York Times (http://www.nytimes.com/2009/06/26/business/media/26adco.html?_r=1).

The 2007 JMLR paper produced a toolkit including a suite of techniques for classification in network data. This toolkit (NetKit-SRL) is available (open source) on sourceforge.net (http://netkit-srl.sourceforge.net/). The paper describes the general problem it solves, the techniques, and provides a systematic evaluation on benchmark data sets.

The final paper in the series, on the on-line brand advertising work (2009), gives a special illustration of design-science work moving from academia toward practice. As such, it needs to take into account additional real-world concerns. For example, marketing/advertising based on social network data can seem creepy. However, this work shows how for on-line advertising social-network-based targeting can be done in a privacy-friendly fashion. This work received an award from the Marketing Science Institute.

Why is this project in the realm of Information Systems (based on item #2 in the award instructions document)? The techniques and systems that are the core of this project are fundamentally information technology based, and the studies address directly improving the productive application of information technology to organizations and their management—in particular, to direct marketing of products and to advertising of brands. The work seems to clearly be Information Systems work (and was done by Information Systems researchers); if there is a question about this, please ask.

Was the project principally led and driven by university based faculty, etc., for R&D or educational purposes? This line of research was principally led by Professor Provost and his current and former students and postdoc. It originated completely as a university-based effort. Next, we partnered with AT&T Research (leading to the 2006 study). Then for the 2009 study the novel development and data collection was done while the entire sub-team was at Media6degrees (where one former student now works). This was necessary in part to get access to the appropriate data for a real evaluation, and also because the development effort required for such a project was (far) beyond our capabilities in the IS group at Stern (please see the paper).

Discuss insights related to design science that have emerged from this specific line of work. Summarize how and why the work has contributed more generally to Information Systems related Design Science in the past or present, or how the work will contribute in the future. Through the various studies, this line of work has shown that the considerable information that is present in the structure of social networks can be used by predictive models built to take advantage of that structure. One theme of insight that runs through the work is the power of the structure itself, rather than complicated algorithms. All four paper show (among other things) that the network structure alone can be remarkably powerful for prediction. The Macskassy & Provost papers reinforce the design science 101 notions that one should carefully choose the right baselines against which to compare, and should conduct lesion studies on a complex artifact, replacing complexity with well-chosen simplicity, to understand when the complexity is warranted and when it is ornamental. The Hill, Provost & Volinsky paper shows that for targeted marketing, the social network structure alone contains very powerful predictive information. For example, in the targeting segments studied, social network neighbors of existing customers were 3 to 5 times more likely to respond to the offer than were customers who did not have a customer as a social network neighbor. In addition, the structure alone being so predictive allowed the identification of likely purchasers who otherwise would have fallen through the cracks, because there was little traditional data about them. The Provost, Dalessandro, Hook, Zhang and Murray paper shows that the structure alone of the constructed quasi-social network can provide very good audience identification for on-line advertising. This means that one need not collect or save any of the traditional targeting information, such as the identity of the browser, demographics, the types of content they visit, etc., allowing for very privacy-friendly on-line advertising.

As mentioned above each of the pre-2009 papers already has received a considerable number of citations. I have not carefully gone through and determined what those authors found citation worthy. The previous paragraph summarizes what is in my view one of the main lines of contribution that runs through this work. The 2009 paper has been in print only since last summer, and I'm hesitant to predict whether it will have any significant impact in the research literature at all, let alone what it will be. However, the focus on designing methods for "privacy-friendly" targeting, as compared to "privacy preserving" data mining, is an important direction and I hope that this work will stimulate further work. The current so-called privacy debate centers on the confidentiality of consumer/customer information in the on-line advertising ecosystem. Most emphasis is placed on two extremes on the confidentiality spectrum: "we can use whatever data we can get our hands on" versus "you can't use my data." Both of these extremes are unacceptable; it is important to recognize both consumers' desire for confidentiality and the benefit of cost efficiency in advertising. Our work provides one point on the design spectrum of "privacy friendly" techniques between these extremes. I hope that this existence proof stimulates more design science work to flesh out this spectrum and give us as consumers and citizens a broader choice in how our data are used.

Supporting documentation:

A Simple Relational Classifier. S. Macskassy and F. Provost. Proceedings of the KDD-2003 Workshop on Multirelational Data Mining. [a preliminary study on non-marketing data that has received surprising attention, with >60 Google Scholar citations]

Network-based Marketing: Identifying likely adopters via consumer networks. S. Hill, F. Provost, and C. Volinsky. ***Statistical Science*** 21 (2) 256–276, 2006.

Classification in Networked Data: A toolkit and a univariate case study. S. Macskassy and F. Provost ***Journal of Machine Learning Research*** 8(May):935--983, 2007.

Audience Selection for On-line Brand Advertising: Privacy-friendly Social Network Targeting. Provost, F., B. Dalessandro, R. Hook, X. Zhang, and A. Murray. In *Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (**KDD 2009**).

The Online Ad That Knows Where Your Friends Shop. **S. Clifford,** New York Times, June 26, 2009.