Computational Social Science 计算社会学

原文：http://www.sciencemag.org/cgi/content/summary/sci;323/5915/721

By David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, Marshall Van Alstyne

翻译：许小可（xiaokeeie@gmail.com）

我们生活在各种网络中。我们定期检查电子邮件，在各处拨打移动电话，刷卡乘坐交通工具，使用信用卡购买商品。在公共场所，可能有监视器来监控我们的行为，在医院，我们的医疗记录以数字形式被保存。我们也很可能写博客给大家看，通过在线社会网络来维护友谊。以上的种种事情都留下了我们的数字脚印，这些踪迹汇聚起来就成为一幅复杂的个人和集体行为图景，同时这些踪迹也有可能改变我们对人生、组织和社会的理解。

虽然收集和分析海量数据的能力已经改变了一些领域如生物学、物理学等，但是数据驱动的“计算社会学”研究却进展缓慢。尽管在经济学、社会学和政治学上的重要期刊都很少关注这一领域，但计算社会学在国际公司如 Google、Yahoo 以及政府部门美国安全局已经开始被研究。计算社会学要么是私人公司和政府部门的专有研究领域；要么虽然某些有特权的研究者使用私有数据发表论文，但这些数据却无法被其他人评价和复制。上述的场景毫无疑问都无助于公众在知识积累、验证和分发上的长期利益。

基于一个开放的学术环境，计算社会学的价值在哪里？能够增强社会对个人和集体行为的理解吗？什么是计算社会学发展的障碍呢？

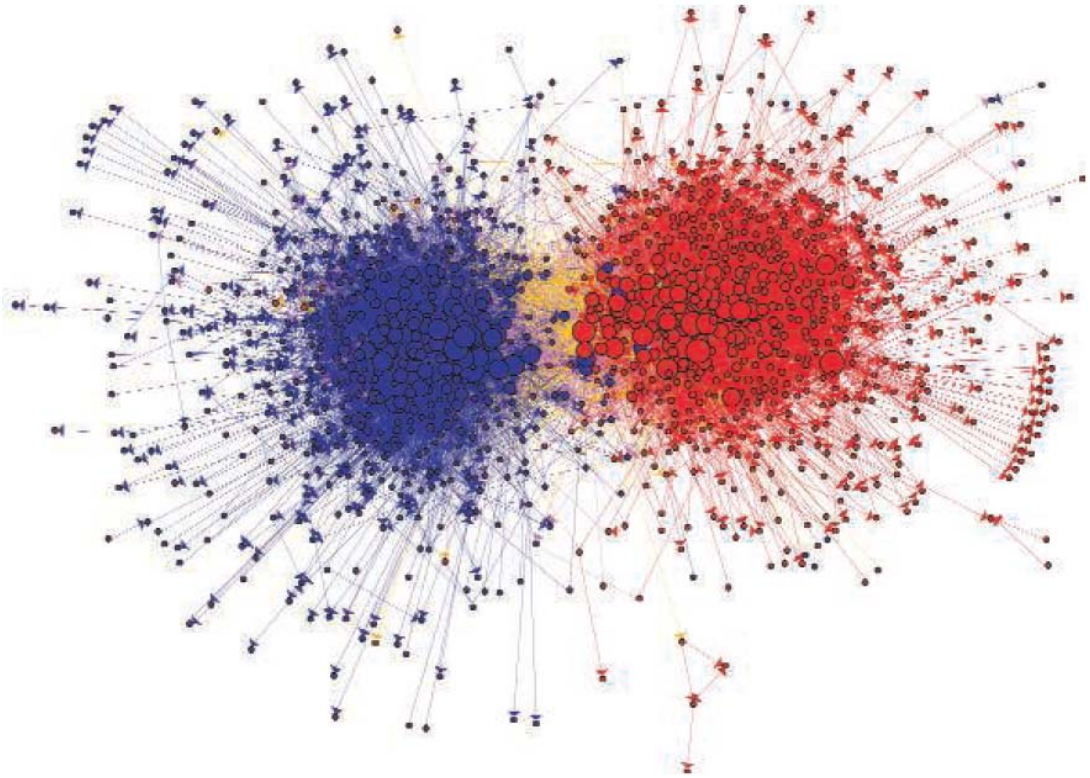
到目前为止，有关人类关系方面的研究主要依赖一次性的、自己报告的数据。新的科技，像视频监控（1），电子邮件和“智能”姓名标记这些手段不仅提供了随着时间发展，在不同时刻的交互关系，而且提供了结构和内容两方面关系信息。例如，团体中的交互关系可以使用电子邮件数据来研究，有关人们交流随时间变化的动态特性等问题也可以被考察：像工作团体是已经稳定下来很少变化，还是他们关系随着时间发生剧烈变化（2）？什么样的交互模式对应着多产的团体和个人（3）？面对面的团体交流能够通过“社会测量法”来评定，而电子设备能够被人戴着从而时刻捕捉人们在物理上的亲密关系、位置、移动以及其他各种个体行为和集体交互等。这些数据有助于解决很多有趣的问题，比如在一个组织内部的亲近关系和交流模式，以及具有杰出表现的个人或集体的信息流模式等（4）。

我们也能够了解社会的“宏观”社会网络信息（5），以及它怎么随着时间进行演化。电话公司拥有数年间他们客户之间通话模式的记录，电子商务门户网站像 Google、Yahoo 等拥有客户相互交流的即时信息数据。这些数据能够描绘社会通信模式的复杂图景吗？这些交互活动中的哪些方式会影响经济生产力或公共健康？不管怎么样，现在追踪人类活动已经变得很简单了（6）。移动电话提供了一种大规模长时间追踪人们移动和物理上是否亲密的方法（7）。这些数据或许会提供有用的流行病学方面的见识：比如一个病原体，像感冒病毒是如何通过物理上的相互接触而在人群中传播的。

互联网提供了一条完全不同的途径来理解人正在说什么以及人们是怎么连接到一起的（8）。例如，在这个刚过去的政治季节中，只要跟踪一下论点、谣言、政治观点以及其他线索在博客空间的传播（9），以及个人在互联网上的“网上冲浪”行为（10），每一个选民究

竟关心什么东西就很清楚了。虚拟世界能自然而然地完全记录每一个人的行为，这也为研究提供了更多的可能性-很多实验在现实中是不可能做和也不被接受的（11）。相似的，社会网络在线站点提供了独特的途径去理解一个人在网络中的地位对整个组织的影响，从他们的感受到他们情绪和健康（12）。自然语言处理已经开始不断增强组织和分析互联网以及其他来源的大量文本材料的能力（13）。

简短地说，计算社会学正在像杠杆一样以前所未有的方式不断增强我们收集和分析数据的宽度、深度和广度。然而，不容易克服的障碍却影响着这一进程。目前存在的方法不能处理数以兆计的时刻变化的整个人类个体之间的交互关系和位置。例如，目前存在的社会网络理论是往往是通过几十个人的一次“快照”得到的数据建立起来的，它怎么能告诉我们有关百万计人口的各种信息之间的相互关系，这些信息包含这些人的位置、商业交易和日常交流等数据。这些大量涌现的人与人之间相互交互的数据能够定量地提供有关人类集体行为的新观点，但是目前我们的研究框架却无法处理这些数据。



从博客空间得到的数据。上图显示的是政治博客社团之间的链接结构（从 2004 年开始）。红色线代表保守派博客，蓝色线代表自由派博客；橙色线代表自由派连向保守派，紫色线表示保守派连向自由派。每个博客的大小反映了其他博客连向它的数量。【经过计算机械协会允许从文献 8 中重画得到】

也有一些制度上的障碍来阻止计算社会学前进。从途径上看，物理和生物学上探索的问题更适应观察和干涉。在发现的过程中，夸克和细胞都不介意我们揭开他们的秘密，也不抗拒我们改变他们的环境。对于基础结构来讲，社会学和计算社会学之间的鸿沟要比生物学和计算生物学之间要大得多，原因主要是计算社会学需要分布式监控，追踪允许以及编码等。这些在社会学中几乎都没有资源可以利用，甚至从物理距离和管理形式上来看，社会学系和工程或计算学系之间的差异要比其他科学之间大得多。

可能最痛苦的挑战是如何保证数据可以获取而又很好保护个人隐私。很多数据都是有所

有权的（如移动电话数据和商业交易信息等）。由 AOL 公司公开它的很多客户“匿名化”搜索记录所造成的大混乱突出了个人或公司通过私人公司分享私人数据的潜在风险（14）。在工业界和学术界之间合作和数据共享的鲁棒模型是必需的，从而来促进研究、保护个人隐私以及为公司提供保护。更一般的讲，恰到好处的处理隐私问题是最基本的。最近美国国家研究委员会有关地理信息系统的报告就特别指出，他们可能会经常性的去掉个人外形特征，并且会仔细地匿名化数据（15）。去年，美国国家健康局和 The Wellcome Trust 突然去掉了一些基因数据库的在线获取功能（16）。这些数据看起来已经匿名化了，仅仅报告了某些基因标记者的总体频率。然而研究表明，在统计上，如果利用数据库中所有个体的全部数据，还是有可能重新确认个体身份的（17）。

因为一条个别的违背保护隐私的小事故就会导致扼杀新生的计算社会学的制度和法律条文产生，所以自我调整的与手续、技术和规则都相关的制度必须要建立起来，从而降低风险，保护潜在的研究。作为自我调整制度的基石，美国机构审查委员会（IRBs）必须增强他们的科技知识来理解入侵和伤害个人的潜在因素，因为新的可能性已经无法用他们当前有关伤害的范例来判断了。很多 IRBs 的人员很难来评估复杂数据被去匿名化的可能性。而且，IRBs 可能需要检查一下是否有必要建立一个专注于保护数据安全的机构。目前，已有的数据在许多组织中传播，这些机构对于数据安全的理解和处理手段是参差不齐的。研究者必须在保留数据做研究的同时开发技术来保护个人隐私。同时，这些系统反过来可能也有助于对于工业界保护客户隐私和数据安全（18）。

最后，计算社会学的发展和其他新兴交叉学科也息息相关（像可持续性发展科学），这就需要发展一个方式来培养新的学者。决定教授职权的委员会和编辑部需要理解和奖赏跨学科发表的努力。最初地，计算社会学需要拥有社会学家和计算学家一起努力。长期地看，这个问题将取决于学术界决定是否应该培养计算社会学学家，或者计量文献社会学家和社会文献计量学家的团队。认知科学的出现为计算社会学的发展提供了一个很好的范例。认知科学涉及的领域包括生物学、哲学和计算科学。它已经吸引了大量资源的投入来创建一个共同领域，而且为过去一代的公共货物作出了很大贡献。我们认为计算社会学具有相似的潜力、值得相似的投入。

References and Notes

1. D. Roy et al., “The Human Speech Project,” Proceedings of the 28th Annual Conference of Cognitive Science Society, Vancouver, BC, Canada, 26 to 29 July 2009.
2. J. P. Eckmann et al. Proc. Natl. Acad. Sci. U.S.A. 101, 14333 (2004).
3. S. Aral, M. Van Alstyne, “Network Structure & Information Advantage,” Proceedings of the Academy of Management Conference, Philadelphia, PA, 3 to 8 August 2007.
4. A. Pentland, Honest Signals: How They Shape Our World(MIT Press, Cambridge, MA, 2008).
5. J.-P. Onnela et al., Proc. Natl. Acad. Sci. U.S.A. 104,7332 (2007).
6. T. Jebara, Y. Song, K. Thadani, “Spectral Clustering and Embedding with Hidden Markov Models,” Proceedings of the European Conference on Machine Learning, Philadelphia, PA, 3 to 6 December 2007.
7. M. C. González et al., Nature 453, 779 (2008).
8. D. Watts, Nature 445, 489 (2007).
9. L. Adamic, N. Glance, in Proceedings of the 3rd International Workshop on Link

Discovery (LINKDD 2005), pp. 36–43; http://doi.acm.org/10.1145/1134271.1134277.
10. J. Teevan, ACM Trans. Inform. Syst. 26, 1 (2008).
11. W. S. Bainbridge, Science 317, 472 (2007).
12. K. Lewis et al., Social Networks 30, 330 (2008).
13. C. Cardie, J. Wilkerson, J. Inf. Technol. Polit. 5, 1 (2008).
14. M. Barbarao, T. Zeller Jr., "A face is exposed for AOL searcher No. 4417749," New York Times, 9 August 2006, p. A1.
15. National Research Council, Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data, M. P. Gutmann, P. Stern, Eds. (National Academy Press, Washington, DC, 2007).
16. J. Felch. "DNA databases blocked from the public," Los Angeles Times, 29 August 2008, p. A31.
17. N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, PLoS Genet. 4, e1000167 (2008).
18. M.V.A. has applied for a patent on an algorithm for protecting privacy of communication content.
19. Additional resources in computational social science can be found in the supporting online material.

Supporting Online Material

www.sciencemag.org/cgi/content/full/323/5915/721/DC1