# 大数据为社会网络分析带来“黄金时代”*

郑　路**

**摘　要**：社会网络分析因为其独特的理论视角和分析方法，在社会学、经济学、政治学、传播学、计算科学、公共健康等领域得到广泛关注和运用。网络社交平台的兴起和随之产生的海量关系数据的出现，为社会网络分析带来了空前的机遇。其一，海量社交数据为解答网络分析中的经典问题提供了更高精度、更完备的数据来源；其二，网络平台使大规模的随机控制实验成为可能，有利于做出更有说服力的因果推断；其三，推动学术界对互联网的社会后果等诸多新的问题的提出和研究。本文通过大量实例就这三个观点展开论述，彰显大数据将社会网络分析带入了“黄金时代”。

**关键词**：社会网络分析；大数据；随机控制实验；研究方法

社会网络分析是近年来在社会科学研究中最令人兴奋的研究领域之一。作为一种研究范式和数据分析方法，社会网络研究从社会学、心理学和人类学中萌芽，经过与数学、物理学和计算机科学的交融，已发展成为一个跨学科的领域，除上述学科之外，在政治学、经济学、传播学、管理科学、公共健康等领域也得到广泛的应用。在过去的10年里，移动互联网及社交媒体的普及为社会网络分析带来了前所未有的机遇和挑战。微信、微博、脸书等网络社交平台催生和记录了人际互动的海量数据，为研究社会网络的重要问题——例如动态网络和因果推论——提供了高质量的数据来源，同时也为运用新的研究方法（例如实验方法）提供了可能。大数据时代的到来对原有的学术传统和方法带来了冲击，在学界中也产生了不少对大数据的质疑。其中最常见的一个问题

*　本研究为国家社科基金重大项目“大数据时代计算社会科学的产生、现状与发展前景研究（16ZDA085）”阶段性成果。

**　郑路，清华大学社会学系副教授、数据科学院 RONG 教授、苏世民书院兼聘教授。

就是：大数据究竟能给我们带来什么新的知识？本文以社会网络分析为例，论述网络社交媒体及大数据至少在以下三个方面推动了该领域的发展：一是能更精确地解答经典问题；二是让大规模的随机控制实验成为可能，从而作出具有说服力的因果推断；三是引发新的研究议题的提出和探索。在对这三个方面进行论述之前，本文先对社会网络分析的基本概念和理论前提进行一个概要性的介绍。

## 一、社会网络分析的独特视角

一般而言，网络是指节点（node）和节点之间的关系（tie）组成的集合。“节点”在社会网分析里一般指行动者，比如个人、群体、组织，也可能是城市或国家，甚至是某些更为抽象的事物，例如文章或专利。“关系”也可能有多种形式，最常见的是人际关系，由此构成朋友网、讨论网、拜年网，也可以是物理空间上的接触，从而构成交通网络、计算机网络等。总之，在更宽泛的层次上，网络的研究也被看成是一切互动形式的研究，从信息交流、情感支持、贸易往来、信用与资金的流动、到疾病的传播、创新与文化要素的扩散，都可以作为网络分析的研究对象。

作为一种独特的研究范式，网络分析有两个隐含的理论前提。一是比起节点自身的特征或属性，节点之间的关系对节点的影响同样重要，在某些时候甚至更大。个体之间在现实世界中存在相互影响是显而易见的常识，但在理论世界中却往往被忽略。例如，经典经济学中的“理性经济人”就被定义为以追求效用最大化为目标的“原子化”个人。同样，作为概率统计中的独立同分布假设（i.i.d），也要求观测值之间是相互独立的。再如，早期的教育学研究通常是用学生的一系列个体属性（例如性别、年龄、教育期望、父母的教育程度、职业、收入等）来解释学业成绩和升学几率；后来受到对小群体和网络研究的影响，研究者发现学生在课外和谁一起玩耍（peer group），以及学校所处社区的社会资本多寡（Cohen，1977；Coleman，1988），对学生的学业成就同样有着重要影响。人际影响的重要性在市场和营销研究里也日益受到关注，研究者发现影响消费倾向和购物选择的因素，不仅仅是消费者的年龄、性别、教育、收入和职业等个人特征，更重要的是他们的朋友或参照群体（reference group）

喜欢什么和买了什么。脸书(Facebook)的市值达到近5 000亿美元[①],并不在于它将人们社会交往的方式从线下搬到了线上,而是因为这个社交平台及其产生的数据对广告投放精准度和销售转化率的提升。脸书的创始人扎克伯格(Zuckerberg)在接受美国《商业周刊》的采访时曾一针见血地指出,"一个广告商可能创作出世界上最有创意的广告,但对可口可乐来说,让你了解到你的朋友们非常爱喝可口可乐,是这家公司能够获得的最有力的支持"(Businessweek, 2010)。我们也可以扪心自问,自己选择用什么品牌的手机或电脑,在多大程度上是受到了周围朋友、同事的影响。

网络分析的另一个理论前提是:节点的行为不仅受到与之直接关联的节点的影响,同时也会受到该节点在整个网络所处的位置以及网络整体结构特征的影响。例如,一个人是处在朋友网的中心或边缘的位置,反映了该个体在群体中的地位和声望高低,影响着个人获取资源或信息的便捷度。对该个体同样重要的,是与她具有朋友关系的其他个体之间,在多大程度上也相互认为对方是朋友。研究表明,如果一个人的好朋友之间并不认为对方是朋友,则会对该个体的心理健康产生负面影响(Bearman and Moody, 2004)。此外,一个网络的密度高低(即节点间两两发生关联的程度),会影响资源或信息在网络中扩散的速度。在社会网络中,密度的高低还反映了这个群体的团结程度(cohesiveness),以及群体信任度的高低,同时还影响到群体规范得以贯彻的可能性。从方法论的角度讲,由于个体有目的的行动总是嵌入在具体的、不断演进的社会关系之中的(Granovettor,1985),能够反映和测量这些社会关系的社会网络,就在微观的个体行动者与宏观的社会制度和文化之间搭建了一个尤其关键的中观层面(meso-level)的联系机制(Granovettor,2017)。

在网络分析的研究中,数据的获取是关键。在网络大数据成为可能之前,绝大多数的网络数据来源于两个渠道,一是以传统的抽样调查为手段,采用"提名法"(name generator and name interpreter)来测量以被访者为中心的个体网的构成。这种方法的优点是能够获得较大规模的样本,但不能得到个体所处的整体网络的特征,从而局限了网络分析的应用。另一个数据来源是对一个规模相对较小的群体(从几十人到几百人)中的所有个体的网络关系进行测量。这种方法使整体网分析成为可能,但缺点是工作量巨大,无法推广到

① 脸书在2017年10月5日的市值为4970亿美元。https://ycharts.com/companies/FB/market_cap.

大规模群体的研究中去。大数据的出现则提供了新的规模巨大的整体网数据来源，研究者可以利用这些数据来解答一些原先因缺乏数据而无法回答的问题。其次，社交网络平台的流行还为利用这些平台开展随机控制实验提供了可能，从而为因果推论提供更有说服力的证据。最后，人们对移动互联网的日益依赖还引发了许多新的、有趣的研究问题。本文接下来从这三个角度，结合三个案例来彰显大数据给社会网络分析带来的机遇。

## 二、利用大数据解答经典问题

“距离”是网络分析中的一个基本概念，指的是连接两个节点的最短路径的长度。经典的“小世界现象”所要回答的问题就是，如果世界上的大多数人都能通过一个朋友或熟人网络联系在一起的话，那么这个网络中的任意两个人之间的平均距离是多少？为解答这个问题，美国心理学家米尔格拉姆(Stanley Milgram)设计了一个实验，把这个问题进行了巧妙的转化：如果让一个住在内布拉斯加州的人把一个包裹通过她的熟人，以及熟人的熟人寄到一个住在马萨诸塞州的陌生人手中，这个包裹需要在人与人之间传递多少次？(Milgram，1967)米尔格拉姆招募了296名实验参与者，告诉他们包裹最终是要寄给一个住在波士顿郊外某个地址的股票经纪人。为了达到这个目的，参与者可以把包裹寄给他们认识的最可能帮助他们达成目标的亲朋好友，并嘱咐对方也遵循同样的办法，以此类推。最终，这296封包裹有64封成功通过熟人链条寄到了那个波士顿的股票经纪人手中，这些成功路径的中位值是6。20多年后，这一研究发现在一部话剧里被演绎成“六度分隔理论”(six degree of separation)。它告诉人们，你不仅可以通过朋友与这个世界相当大的一部分人有联系，而且你和这些陌生人之间的距离短得惊人。这也难怪当我们在社交场合发现和初次见面的某人有共同认识的朋友时，都会由衷地感叹一句“世界好小啊！”

然而，从上面对实验过程的描述，我们不难看到“六度分隔”的结论是有很大问题的。首先，这个数值是仅仅基于送达的64个包裹而得出的结论，其他230多个未成功的信息则被忽略掉了，因此六度分隔可能被低估了。其次，即使是这些成功送达的包裹，我们也无法知道它们经过的路径是否是最短的，受此影响，六度分隔又有被高估的可能。尽管社会学家邓肯.瓦茨(Duncan Watts)和他的导师史蒂文·斯托加茨(Steven Strogatz)于1998年在《自然》杂

志上证明了六度分隔的数学原理(Watts & Strogatz, 1998),对小世界现象痴迷的瓦茨在2002年和他的两个合作者再做了一次米尔格拉姆的实验。这次是将参与者扩大到全球,用电子邮件替代了原实验中的包裹(Dodds, Muhamad and Watts,2003)。研究者在全球招募了6.1万名电子邮件使用者,要求他们通过熟人,以电邮为媒介联系到在全球范围内随机抽取的某个目标人。这个实验得到的结果也再度证实了六度分隔的结论。比照米尔格拉姆最初的实验设计,瓦茨的再次尝试同样无法克服上面提到的两大缺陷,直到大数据的出现。

2006年,两位研究者获得当时使用微软实时通信软件的2.4亿活跃用户的数据,他们将在一月之内有过对话的两个账户之间定义为熟人关系,从而获得了一个熟人关系的整体网络。这个新的数据源克服了信息缺失的问题,同时也能甄别出两个用户之间的最短路径,从而计算出了用户之间6.6的平均距离。而脸书公司在2012年发表的一份研究报告,则告诉世人脸书用户的平均距离为4.74,并且随着脸书用户量的增加,用户间的平均距离还有缩小的趋势(Backstrom et al. 2012)。从这个例子可以看到,网络大数据的出现让我们对"小世界现象"这一经典问题的研究和深入成为可能。

## 三、社交平台使大规模的随机控制实验成为可能

在社会科学的因果推断中,随机控制实验是黄金标准。尤其是当样本量足够大时,随机误差的可能性大大降低,随机化的过程保证了除自变量(也称处理变量,treatment variable)之外的其他变量不会对因变量造成影响,从而建立起自变量和因变量之间的某种因果性联系。2012年发表在《自然》杂志的一项堪称经典的文章里,研究者采用了随机控制实验的研究设计来检验和测量人际网络对投票行为的影响程度(Bond e. al.,2012)。在2010年11月2日美国国会选举投票日,研究者与脸书公司合作,将当天使用了脸书的、身处美国、并达到法定投票年龄(18岁)的6 100多万用户,根据其是否接收到与选举相关的信息及何种信息,分成了三个组别:即"社交消息(social message)"组、"资讯消息(informational message)"组或控制组。被随机划分到控制组用户的"新闻推送(News Feed)"栏中不会出现与选举相关的任何信息;"资讯消息"组的用户会在新闻推送栏的顶部收到一则鼓励他们投票的信息,同时还包括了一个指向本地投票站地理信息的链接,一个报告"我投票了"的按钮,以

及已有多少脸书用户报告已投过票的数字；“社交信息”组的用户在“资讯消息”组用户收到的上述信息之外，额外增加了她的脸书朋友里报告已经投过票的6位朋友的头像照片。研究者关注的是这三组用户在三个方面的行为数据，一是点击“我投票了”的按钮，二是点击提供投票站地理位置的链接，三是经过线下数据核实的该脸书用户是否真正投过票的信息。

分析发现，和“资讯消息”组的用户相比，“社交信息”组的用户在看到报告已经投票的朋友的照片后，按“我投票了”按钮的概率会增加2.08%。“社交信息”组的用户点击投票点信息的概率也比前者高了0.26%。这表明人际影响比单纯的信息更能够引发人们的投票倾向。在与公开的投票纪录进行对照之后，发现接收到“社交信息”的用户的实际投票率比未接受到任何信息的控制组用户的投票率高0.39%，比接收到“资讯信息”的用户的投票率也高出0.39%。换句话讲，资讯信息在鼓励大家实际投票的行为上，并没有任何影响，而了解到你的脸书朋友报告投票了，则会对个人的政治自我表达、信息搜集和真实世界的投票行为产生显著的影响。据估计，脸书上的“社交信息”直接增加了6万投票人，间接增加了28万投票人，总计增加了34万选票。进一步的分析还根据用户间互动的频率对强关系和弱关系进行了区分，发现强关系的朋友虽然仅占脸书里所有朋友关系的7%，但几乎所有增加的投票行为都只能在强关系中才能得以实现。这进一步验证了强关系传递影响、弱关系传递信息的假设。这一涉及6 000万脸书用户的空前巨大规模的随机控制实验不仅使人信服地验证了线上社交信息对于政治参与的因果效用，而且提供了在社交平台上运用实验方法以开展社会科学研究的典范。

## 四、大数据激发对新问题的研究

当今世界，政治价值倾向的极端化日益突出，引发社会的分裂与动荡。研究者将2004年美国总统大选之前政治性博客的网络结构进行了分析，展现了一个两极化的政治博客空间（Adamic and Glance，2005）。支持共和党候选人的博客和支持民主党候选人的博客之间鲜有对话和交流，更多的是本阵营内部的自说自话，两大阵营甚至在关注和讨论的新闻和议题上也大相径庭。虽然政治倾向两极化趋势的背后有深刻的政治和社会原因，人们也不禁在思考上述对网络空间的研究是反映了美国政治的两极化，还是网络工具（尤其是推荐系统）的存在也促进了这一过程的发展。换句话说，互联网的存在究竟是

给公众提供了更丰富、更多元的观点和信息,从而使人们变得更加开放和宽容,还是让公众更容易找到和接触具有相同观念的人或信息,从而使他们的已有观念更容易被强化,变得更加的固执和保守?尤其是互联网公司普遍使用了基于用户行为偏好分析的推荐系统,这是否会强化人们所接触信息的同质化和单一化?

为了回答这一问题,笔者与搜狐新闻客户端合作,分析了该公司1亿用户历时半年的阅读行为数据(Zhang, Zheng and Peng, 2017)。我们的研究问题是,在推荐系统的作用下,随着时间的推移,用户阅读的新闻主题是更加多元化,还是相反?分析发现,在控制了总体内容多样性(用新闻标题的词法共现网的传导性来测量)的条件下,新闻客户端用户阅读种类的多样性随时间在降低。这一发现与《科学》杂志2015年发表的一项对脸书用户的研究相互呼应,该研究发现,在算法排序的作用下,脸书用户所接触到的内容在意识形态维度的多样性在逐渐降低(Bakshy, Messing and Adamic 2015)。因此,不论是社交媒体,还是网络新闻媒体,推荐系统的使用都会起到降低用户阅读内容多样化的作用。我们的研究还对不同性别进行了比较,发现与女性用户相比,男性用户阅读内容多样性随时间降低的趋势较女性小。对性别差异(以及社会阶层差异)的解释,还有待于进一步的研究。我们的研究揭示,对互联网公司而言,完全按照用户偏好来设计推荐系统,从长远来看反而会减少用户使用该应用阅读新闻的总量,从而有损其商业利益。对整个社会而言,新闻推荐系统导致的自我强化机制则可能导致新闻受众的碎片化和价值取向的极端化。这个例子说明大数据的时代不仅给研究者提供了更高精度和准确度的数据,还促发了我们对新产生的研究问题的思考和探索。

## 五、结论

从上文的介绍和论述可以看到,社会网络分析有其独特的理论视角和研究方法,它与方兴未艾的网络社交媒体具有天然的亲和力。后者产生的海量数据以及作为潜在的研究平台,不仅将促进社会网络分析的进一步发展,而且对推动计算社会科学、甚至整个社会科学领域的研究都具有不可估量的作用。这一序幕才刚刚拉开,我们窥见的只是精彩的开始。

**参考文献**

Adamic，L，and N Glance. 2005. The Political Blogosphere and the 2004 U. S. Election：Divided They Blog. *Proceedings of the 3rd International Workshop on Link Discovery* 36－43.

Backstrom，L，P Boldi，M Rosa. et al. 2012. Four Degree of Separation. Facebook Data Science Team. http://arxiv.org/abs/1111.4570.

Bakshy，E，SMessing and L Adamic. 2015. Exposure to Ideologically Diverse News and Opinion on Facebook. *Science* 348(6239)：1130－1132.

Bearman，PS，and J Moody. 2004. Suicide and Friendships Among American Adolescents. *American Journal of Public Health* 94(1)：89－95.

Bond，R，C Fariss，J Jone，et al. 2012. A 61-million-person Experiment in Social Influence and Political Mobilization.*Nature* 489(7415)：295－298.

Businessweek. 2010. Facebook Sells Your Friends.

Cohen，J. 1977. Sources of Peer Group Homogeneity. *Sociology of Education* 50：227－241.

Coleman，J. 1988. Social Capital in the Creation of Human Capital. *American Journal of Sociology* 94(1)：95－120.

Dodds，P，R Muhamad and D Watts. 2003. An Experimental Study of Search in Global Social Networks. *Science* 301(5634)：827－829.

Granovetter，M. 1985. Economic Action and Social Structure：The Problem of Embeddedness. *American Journal of Sociology* 91：481－510.

Granovetter，M. 2017. *Society and Economy：Framework and Principles*. Belknap Press of Harvard University Press：Cambridge，Massachusetts.

Leskovec，J，and E Horvitz.2008.Worldwide Buzz：Planetary-scale Views on an Instant- messaging Network. *Proceedings of 17th International World Wide Web Conference*.

Milgram，S.1967.The Small-world Problem. *Psychology Today* 2：60－67.

Watts，D J，and S H Strogatz. 1998. Collective Dynamics of“Small-world” Networks.*Nature* 393(6684)：440－442.

Zhang，L，L Zheng，and T Peng. 2017. Structurally Embedded News Consumption on Mobile News Applications. *Information Processing and Management* 53(5)：1242－1253.

# Big Data Brings Social Network Analysis into A Golden Age

Zheng Lu

**Abstract**: Distinguished by its unique theoretical perspective and analytic methods, social network analysis (SNA) has gained much traction in multiple disciplines, such as sociology, economics, political science, communication, computer science, and public health, etc. The emergence of social media platforms and the resulting big data have brougtht unprecedented resources and opportunities for SNA. To begin with, voluminous relational data are often of high precision and completeness, enabling researchers to reexamine many classic yet unsolved research questions. Secondly, social media platforms make large scale randomized controlled experiments possible, providing more convincing evidence for causal inferences. Thirdly, big data push researcher to ask and explore new questions brought by the ubiquitous use of Internet and related technologies. The paper draws on various cases to illustrate above three aspects, highlighting the incoming of a Golden Age for SNA thanks to big data.

**Key words**: social network analysis (SNA); big data; randomized controlled experiment; research method