

论社会学理论导引的大数据研究*

——大数据、理论与预测模型的三角对话

罗家德 刘济帆 杨鲲昊 傅晓明

提要:计算社会科学把社会科学理论以及研究方法与大数据分析融为一体,一方面为大数据分析开启了很多新议题,理论指导下的定性、定量调查也可以为数据挖掘的结果提供校准的扎根真相;另一方面,在大数据挖掘的结果中可以找到建构理论的线索,提供验证理论的资料,进而指导预测模型的建构,推论并解释更多的现象。本文以中国风险投资产业网络数据为例,展示了数据挖掘、社会学理论与预测模型间的三角对话,进一步呈现了以理论导引的大数据分析的方法论。

关键词:大数据 计算社会科学 动态网 圈子理论 嵌入理论

一、计算社会科学方法论

大数据^①的出现使计算社会科学(computational social science)得到了极大的关注,然而早期的大数据研究聚焦在应用上,因而只是把收集到的数据当母体,不强调随机抽样,而且主要用于描述性统计和相关分析,不重视因果推论(causal inference)。这样以数据挖掘(data mining)为主的大数据研究方法(Viktor & Kenneth, 2014)往往回答了“是何”(what),而不能回答“如何”(how,即了解过程发展的机制)以及“为何”(why,即得到因果关系)。少了“如何”与“为何”的回答,从

* 本文展示的研究范例来自作者与他人合作的多篇文章(罗家德等,2014;罗家德、曹立坤、郭戎,2018;罗家德、樊瑛、郭晴、周建林、刘济帆、李睿琪,2018;Zhou et al., 2016; Zhou et al., 2017; Luo et al., 2017, 2018; Gu et al., 2017),在此向所有的作者表示感谢。此外,感谢清华大学数据科学院支持的清华大学社科学院与德国哥廷根大学学生交换项目(IDS-SSP-2017001),同时感谢腾讯研究项目“以微信及QQ大数据分析个人人脉”的资金支持(20162001703)。

① 本文所指涉的“大数据”是相对于结构化数据(structural data,即我们一般经过社会调查与整理二手资料得出各种变量的数据库)而言的,是行动主体在网上活动时留下的电子足迹的非结构化数据,请参考Fu et al., 2017。

相关研究中得到的预测模型便欠缺了因果推论的能力(Rubin, 1974)。大数据分析最常被引用的例子就是超市收银数据显示买尿布和买啤酒高度相关,因而建议在尿布区旁放上啤酒。但我们应当追问这些消费者是怎样的一群人、他们的消费风格是什么以及他们做出消费决策的心理是什么,有了回答这些问题的理论之后,预测模型才能知道推论的范围。比如,当下有效的预测什么时候会失效?在美国有效的预测在中国是否有效?这样的购买行为还可以推论到其他什么商品?

通过大数据的数据挖掘得到的预测因子(predictors)以及行为模式(behavioral patterns)可以证伪过去不同预测背后的理论,却无法自证新理论的成立(Popper, 1965)。针对由资料归纳得到的结果,我们仍需要进行诠释,形成理论,发展假设,收集相关学术社群能共同接受的“事实”,^①与竞争理论的假设进行对话,共同接受“事实”的验证,从而得到学术社群进一步的认可(Lakatos, 1980)。使用理论建立预测模型,才可以推论到不同时间,不同范畴以及不同地区、文化的新“事实”,进而建构出可以推论的预测模型(Galison, 1987)。简言之,能够作出推论的是理论,而不是数据本身以及数据挖掘的结果。所以,社会科学理论与数据挖掘的对话对预测模型的推论能力至关重要。

计算社会科学的发展,正是把社会科学的理论带入数据挖掘之中,尤其是大数据的数据挖掘之中。一方面,可以使用大数据对理论发展出来的假设进行验证,梅西(Michael Macy)以大范围地区内电话通话频率来衡量社区社会资本,证实社区社会资本能影响社区经济发展(Eagle et al., 2010)。同时,在理论不够明确时,数据挖掘也能给理论提供启发,我们可以诠释(interpretation)数据挖掘的结果,与可能解释此一现象的理论展开对话,并在此过程中发展新理论。

另一方面,理论反过来可以指导数据挖掘的方向,比如邓巴(R. Dunbar)主张人脉会因亲疏远近的不同而分成几个圈层,并使用社交网络的资料来加以分析(Dunbar et al., 2015),由此挖掘出区隔亲疏圈层的算法。另外,理论指导下的定性研究与定量调查可以用来收集资料,以校正数据挖掘得到的结果。比如,科辛斯基(M. W. Kosinski)利用脸书大数据去计算脸书使用者的大五人格(Kosinski et al., 2016)时,

^① “事实”加上引号,是为了避免陷入是否有客观存在事实的争论中,而以交互主观(inter-subjective)承认的资料作为验证竞争性理论的依据。

先以调查方法在现实中收集到一群人的人格测量结果,再在脸书上将这些人的网上行为记录下来,这样收集到的现实“事实”可以验证数据挖掘结果的有效性(Kosinski et al, 2016)。

针对数据挖掘的目标现象,在社会科学理论与方法指导下以定性、定量调查得到的资料也被称为扎根真相(ground truth)。扎根真相原来是遥感学界的用语(Seager, 1995),用来指称高空或卫星成像后,在分析中人们想知道在地面上(ground)其所摄影的真实物件到底是什么(truth)。此概念用于数据挖掘过程中,则指涉的是挖掘出来的预测因子或行为模式在现实中到底存在不存在,以及挖掘出来的目标现象和真实世界中的“事实”到底有多少差别。换言之,理论指导下的调查可以提供检验数据挖掘结果的扎根真相。

一个理论一旦获得了“证实”,^①我们就可以使用理论来建立预测模型,预测模型不仅可以在一定的准确率上还原原有的资料,而且可以在理论演绎中推论出新的“事实”。若预测模型还原原有资料的准确率还有提高的空间,或是它所预测的“新事实”与实际收集到的资料有一定差别,都表示理论还有改善的空间,因此研究者又启动一轮数据挖掘与理论间的对话,从新数据挖掘中得到启发,诠释挖掘的结果并与可能的相关理论对话,对原有理论进行修正,并以资料(大数据、调查数据或二手数据)加以验证。同样,经过修正的理论又可以提出新的预测模型,推论新的“事实”,当然也可能引发新一轮的数据挖掘。

本文将以上所述的研究过程制作成图1。简言之,理论对大数据分析的贡献在于提供了丰富的新议题,指明了可以研究的新方向,比如大五人格、人脉邓巴圈(Dunbar Circle)以及社区社会资本等,同时在理论指导下用定性、定量方法收集到的“事实”可以作为数据挖掘的扎根真相,提高挖掘成果的准确率。而大数据除了可以用来当作验证理论的资料外,其挖掘成果也可以通过诠释和与其他理论对话来加以演绎,得到新的理论或修正旧的理论。理论又进一步指导预测模型的建构,而预测模型又推论出新的“事实”,无论是在时间上、文化环境上还是

^① 这里“证实”再度加上引号,指的是在相关学术社群可以接受的“事实”面前,该理论在相关竞争理论中具有更好的解释力,这样的理论竞争以学术社群共同认可的“事实”为依据进行竞争,从而避免了逻辑实证论指称的理论实证,在事实面前理论提出的假设在一定显著水平上能够成立,从而理论得到确证。请参考上一页脚注^①以及 Hempel, 1966。

新的范畴上,新“事实”又会有相应的新数据,如此周而复始,使理论不断得以修正,也使推论扩展到更广阔的领域中。

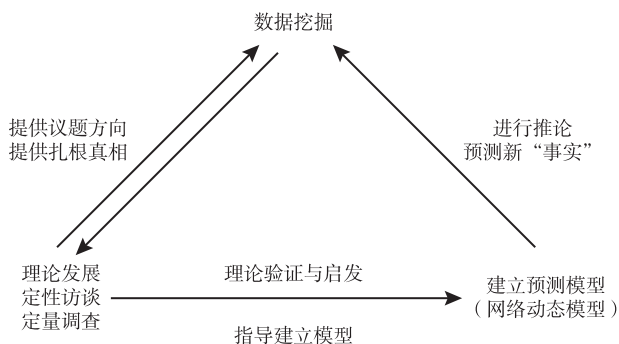


图1 数据挖掘、理论与预测模型间的三角对话

换言之,大数据本身与数据挖掘无法进行推论,其得到的相关结果只能在有限的时空中做应用型的预测。是理论的演绎帮助我们进行更广阔的推论,从而在一轮又一轮如图1的三角对话中扩展大数据分析的方向,也不断地修正社会科学的理论,得到预测模型,推论出更广领域中的现象。下面笔者就以一组风险投资企业的联合投资网络资料^①作范例,将此三角对话过程加以完整的展示。

二、大数据对理论发展的影响

在理论导引的大数据研究中有一个重要的步骤就是将网上电子印迹数据转化成理论发展所需的变量。非结构化数据(unstructural data)并非天然适用于理论演绎和验证,对于上述三角对话中的动态网模拟等更为复杂的方法来说,更是需要精细、有效的整理和结构化。大数据与理论有效对话往往要经过一个阶段,即通过一套算法变成结构化的

^① 此一资料的原始来源是网上收集到的风险投资公司公布的投资信息,如果两家风险投资公司(私募基金与天使投资排除在外)宣布在同一时点上投资了同一家企业,则视为有一次联合投资。原始数据经过清科集团研究中心(私募通数据库)初步整理,本研究团队在网上收集必要信息加以补充,得到联合投资网络。

数据,才能为理论所用。比如,我们在分析中国社会的阶层分布时,需要先对每个人的社经地位属性(Social Economic Status,简称 SES)有所了解,而过去社会学的长期研究得到的定量指标包括个人的教育、收入、财富、职业、职级等资料;进一步而言,家庭背景也很重要,最好还有父亲、母亲的职业与教育,然后综合建模可以得到 SES 的衡量分数。但这些资料都无法在网上的电子印迹数据中找到,所以我们会去找针对 SES 研究的调查(survey)数据以得到扎根真相,然后在大数据中找可能的行为结果,比如什么阶层的人会住在哪里、买哪些东西等,从而收集其移动轨迹中晚上的常驻地(最可能的住家所在地)、白天的常驻地(最可能的办公所在地),以及该地地价、个人浏览什么网页(侧面了解一个人的消费风格)、网上下单买了什么商品(侧面了解一个人的收入状况)等电子印迹中能找到的数据,以这些数据得出一套算法,预测这个人的 SES。根据所得的扎根真相的数量,比如该调查只访问了 20000 人,那么研究者会在全网中找出这 20000 人,寻找其算法,并将大数据转化为结构化数据,从而做出更多的理论验证与理论推导。这时,数据量受限于扎根真相,可能有大有小,其大小自然对预测精确度有影响。当得到一套可以将大数据转化为理论上有用的指标的算法后,甚或进一步以理论推论出一套行为模式后,在理论认可的范围内,可以将算法与行为模式应用于所有网民,这时使用的数据量会变得极为庞大。

简言之,以理论驱动的大数据分析其数据来源体量非常大,经过一套算法得到和理论匹配的结构化数据时,数据量可能大也可能小,但推论到更多人、更长时间和更广范畴时,又会使用体量巨大的数据。比如当我们能匹配出 50000 人的 SES 分数和他们的网上行为轨迹,并得到令我们满意的预测精度时,就可以用以推论类似这 50000 人(如果这 50000 人不是随机抽样取得,而属于某一社会类属,则推论止于此一类属)的所有网民的 SES,并以推论得到的可能数以千万计的数据进行进一步的理论分析。

对于任何一个社会科学领域的大数据研究,这种结构化的大数据都是必要的。在上文提到的邓巴圈研究中,不论是对于脸书数据,还是微信、人人网等社交软件数据,或移动、联通等电话通信数据,研究者首

先要解决如何从每天以 TB 级^①增长的数据中找到特定的用户,提取该用户与其他人的互动记录并构建互动网络。这种处理方法完全依赖大数据技术,不仅匹配计算过程需要使用分布式算法,计算结果的存储也需要放在分布式数据库中,这不同于经典关系型数据库的数据结构。此外,对于非公开数据的处理以及拟合全网络度分布过程也相当复杂。当这些过程结束之后,研究者最终得到了一笔和扎根真相配套的数据以及最合适的预测因子(predictors),并由此导引出一套预测最精确的算法,整合出结构化的人脉数据。

本文需要强调的是,尽管社会科学领域的大数据研究中使用的数据很可能看起来与经典社会科学研究,尤其是重视计量的社会科学所使用的结构化数据没有太大的区别,但实际上,这类数据本身的数据规模、获取过程,以及最为重要的,这类数据所能支持的研究议题本身都是不同的。而结构化的大数据所能支持的最为重要研究议题之一就是动态复杂网络研究。

本文接下来以研究风险投资公司形成的联合投资网络的结构为例来加以说明。该研究具体的内容会在下文进行详细介绍。如果单看数据清理的部分,在经济和金融系统高度数字化的当下,风险投资公司的投资数据对于研究者来说其实非常容易找到,大量上市公司的财报、经济新闻、风险投资公司的公开资料中都含有丰富和详细的投资事件信息。但使用这些信息构建联合投资网络却并不容易,这正与大数据的非结构化特点有关。即使经过网络爬虫算法和关系型数据库技术初步整理后的投资数据往往也是大量零散的投资事件,这也是本文所介绍的风险投资公司研究中研究者面对的实际情况。^② 如何匹配这些投资事件,形成投资网络,这本身就是一个问题。尤其对于大量存在的信息不全的投资事件来说,任何一个社会科学研究者都了解缺失值的处理对于结论的合理性有多么重要,而非结构化数据中的缺失值的存在方式决定了普遍通用的缺失值处理技术并不能完全适用。

对于一项投资时间缺失的投资事件,研究者首先需要依赖该时间的其他关键信息,如投资额、币种、交易发生地、被投资方信息,在网上进

① TB 系数据规模度量单位,在计算机理论中 1TB = 1024GB,在实际存储和硬盘生产中,1TB = 1000GB。

② 即清科集团研究中心的私募通数据库。

行搜索,找到匹配的投资事件,并且根据网络上该条信息留下的时间戳或者新闻报道的时间,补全在原数据库中缺失的时间信息。与之类似,对于投资方缺失的非公开投资事件,研究者如果简单地删去或者将缺失值赋值为同一值,都将导致在最终的网络中出现极为偏颇的度分布和自指向型节点,这会导致下文所介绍部分指标无法计算。因此研究者必须使用关系型数据库进行多重比对,以区分不同的非公开投资事件。

由于大数据特点决定了缺失值本身也是大量的,因此以上提到的这些步骤都需要通过算法实现自动处理。研究者之所以不能采用例如多重插补法等经典的缺失值处理方法,其本质原因就在于此时的大数据本身是没有结构化的。经典的社会科学数理方法也许并不比时兴的大数据方法简单,但依赖于已经生成了明确变量的结构化数据集。而在上述例子中可以很清楚地看到,研究者面对的大数据本质上并不直接包含研究需要的变量(联合投资网络及对应的网络指标),因此需要用大量新方法进行处理,将其结构化。这种结构化过程远比原来社会科学方法中的数据清理过程复杂。再以数据匹配这一看似非常简单的过程为例,在风险投资公司的原始数据中,往往公司名会同时以全称和简称的形式分别存储于不同数据库中,需要一个将两者识别和匹配在一起的匹配过程。但是简称和全称的匹配实际上并不存在固定的规则,比方说并非所有公司都是以全称的前两个字作为简称,因此仅是匹配数据这个步骤也需要使用自然语言处理、分词、正则表达式等多项技术。而结构化的结果则可以得到本文所说的结构化的大数据,即能够用于理论演绎和因果推论、包含关键理论变量而又保留了互联网时代个人电子足迹的数据量、足以支持动态网方法的数据。下文介绍的主要研究中使用的数据,本质上都是这种从大量电子印迹数据中整合得来的结构化数据。

因此,在展示上文三角对话的过程中,本文将以动态网模拟作为预测模型的范例,一方面是因为社会科学理论指导了数据挖掘的进行,给大数据分析提供了更多的方向;另一方面,大数据不仅可以用来检验理论,为理论建构提供启发,同时也拓展了理论建构的新方向,尤其是对动态复杂系统理论的建构与检验。

动态复杂社会系统的演化一定是个人行为与整体社会网结构的共同演化(Padgett & Powell, 2012),过去大范围、长时段的多次资料收集

十分困难,而非结构化(unstructural data)电子足迹的数据正好弥补了这样的不足。以网络结构演化为例,过去的个人中心社会网(ego-centered network)资料的收集固然可以涉及较大范围的随机抽样的资料,但收集到的却都是个人的社会网情况,不足以探究较大范围的整体网络结构。整体网(whole network)资料固然可用于分析一定范围内网络的整体结构,但过去的调查方法使这个范围被限定在很小的区间内,收集一个数百人的整体网资料就变得十分困难,枉论数以万计、百万计的大型社会系统(Wasserman & Faust, 1994)。网络动态的资料收集就更困难了,相同的人在不同时段中被问及私密的个人交往情况,两三次就产生了戒心,因此一些通过调查得到的三期、五期比较静态资料已属难能可贵(Burt & Burzynska, 2017),枉论动态网资料了。

社交网络网站和软件的出现,如脸书、推特、QQ和微信等,使此一情况完全改观,几百万人、几千万人的人际关系网络可以被记录下来,而且这样的记录已存在了十多年,通过整理一月一期或一季一期的电子足迹,就可以取得数十甚至上百期网络结构演化的数据。所以大数据的出现使得过去网络动态学(network dynamics)理论建构与理论假设的检验从近乎不可能变为可能。

为什么网络动态学理论的建构及其所指导的预测模型如此重要?

无论是在自然科学还是社会科学中,复杂理论(complex theory)的出现都是为了纠正过去的理论中化约主义(reductionism)的倾向(Prigogine, 1955)。在社会科学中最著名的论述就是格兰诺维特(M. Granovetter)提出的“低度社会化”(under-socialization)与“过度社会化”(over-socialization)的问题(Granovetter, 1985)。前者指的是个人行为的线性加总就是集体行为,集体取决于个体;后者指的是集体形成的力量形塑了个体,个体取决于集体。但其实两者都预设了原子化的个人,犯了化约主义的过度简化的错误,所以都忽略了集体不是个体的简单加总,个体会联结在一起,形成大规模的复杂社会网络,在个体行动与社会网的共同作用下才产生了集体行动(Granovetter, 2017)。

科尔曼(Coleman, 1990)表达了十分相似的论点。如图2所示,按化约主义的观点,都是集体因素解释集体结果(“4”的过程)、个体因素解释个体结果(“2”的过程),由集体到个体时(“1”的过程)就是集体决定个体的“过度社会化”观点,由个体到集体时(“3”的过程)则是个体行为的加总就是集体行为的“低度社会化”观点。科尔曼认为这样

的解释缺少了人际间的互动、关系、社会网络和网络形成的结构。从社会网的观点观之,“1”的过程包括了四类研究,集体的力量可以被视为场力,包括了信息类与规范类的场力(DiMaggio & Powell, 1982)。第一类研究认为它会影响个人的关系以及人脉网络的形成,第二类研究则认为这些关系与个人中心社会网,无论是因为朋友间的交互影响还是个人因社会网而取得的社会资本(Lin, 2001),都会影响个人的行为结果。第三类研究认为场力也会影响个人周围较大网络的变化,从而使个人在网络中的结构位置发生变化。第四类研究认为,个人的结构位置,比如结构洞(Burt, 1992)或封闭网络的中心性(Prigogine, 1955)也会影响个人的行为结果。

“3”的过程则包含了三类研究:一是个人切断关系、建立关系的行为会造成一个网络结构较大的变化(Powell et al., 2005),这正是网络动态学探讨的议题;二是网络结构与人们行为的演化会涌现(emergence)出集体行动(Padgget & Powell, 2012);三是长期、持续、范围较大又有影响力的集体行动终于形成了新的场力,成为“1”的过程中形塑个人关系与结构位置的力量(罗家德等,2008)。

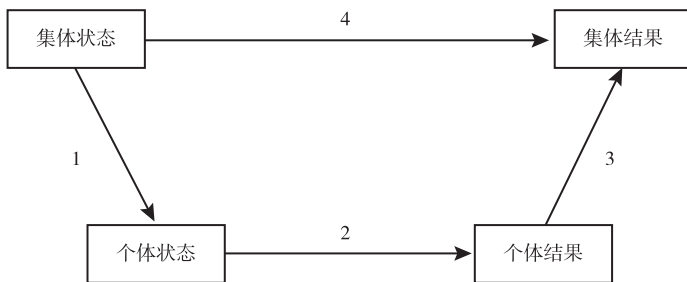


图2 科尔曼论点示意图

综上所述,我们可以看到,大数据的贡献最重要的焦点正好是“3”的过程。网络动态学的研究以及结构与行为共同演化涌现出集体行动的研究,过去只能有一些理论的臆测,验证理论的资料却很少,自然难以更深入、更细致地不断发展与修正理论。大数据的出现使图2中的研究闭环得以完整,进而可以分析大规模复杂网的动态变化、重大集体行动的涌现(如重大创新、社会运动、革命爆发等)、复杂社会系统的非常态演化(如金融风暴、景气突转、社会变迁),以及经济、社会、政治体

制转型等非线性的发展。

简言之,如前所述,社会科学理论一方面可以指导大数据分析的发展,寻找新的议题,监督数据挖掘的结果,进行更广泛的推论;另一方面,也因大数据的加入而有了更大的发展空间,过去一些很难被解释和验证的议题如今有了可能,成为理论发展的新蓝海。

一个产业的景气为什么,以及在何时会突然逆转?拐点何时到来?有无指标可寻?这一直是我們感兴趣的议题,这类议题正是复杂理论想回答的。但在回答这问题之前,我们除了要问产业中的行动者行为有何变化之外,也一定要问这产业的网络结构是怎样的、结构如何演化。下面本文就以风险投资(Venture Capital, VC)产业(以下简称风投产业)为例,思考风投产业网的结构分析问题。

三、理论指导的扎根真相

当我们想了解风险投资公司形成的联合投资网络的结构时,大数据分析者往往会用社群侦测的技术(community detection),将整个产业网分成数个到十余个社群,但这样分出来的社群意义何在?每一个社群代表怎样的投资者?他们的投资行为是怎样的?为何会抱团在一块?比如,当风投产业网中的上千家公司分成了十群时,我们不禁要问,真实世界真的可以如此区分吗?每一群代表什么意义?为什么这些风投企业会抱团聚集?它们会如何演化?对于这些问题,圈子理论提供了一个理论指导的方向(Luo, 2016)。

在中国人日常生活中,圈子或称小圈子通常被定义为职场中的个人中心人脉网——“是从自我中心社会网发展而来的,往往有一个中心人物(或一小组中心人物),只包括他(或他们)的拟似家人和熟人这样的强连带”(罗家德,2012)。^①在风投产业中,圈子则指的是一个VC公司的频繁合作的伙伴形成的合作网。这是中国社会中人们建立差序格局人脉网(费孝通,1998)在职场上的反映,这些频繁合作的VC的一般合伙人之间往往或多或少保持着个人间的交情。另外,在每一次的

^① 杨国枢将中国人的差序格局人脉圈分成家人、熟人与生人三层,一般而言,前两者是强连带,后者是弱连带。请参考杨国枢,1993。

投资中都有一个主投者,主投者会负责投资计划的拟定,并且会在被投企业的董事会中占有席位,而其他的投资人都是跟随者。在一个投资案的最初几轮中,往往是主投者邀请,跟随者才得以加入联合投资之中。因此,一个经常主投的投资者的身旁就聚集了一群或亲或疏的追随者,其中那些较亲密且合作较频繁的追随者就形成了以该主投者为中心的小圈子(罗家德等,2014)。

虽然每一个投资者都想建立自己的圈子,不少投资者也有着或大或小的圈子,然而只有那些规模大、投资频繁或具有行业专门知识等特殊资源的 VC 会经常主导投资,他们形成的圈子也才在整个风投产业中具有影响力,因此我们称之为“产业领袖”。找出产业领袖以及他们的圈子,观察他们之间的合作、竞争以及圈子的演化,是我们了解整个产业网结构的一大关键(罗家德等,2018b)。所以,如何找出一套可以计算出产业网中“产业领袖”的算法是一个关键的手点。

这个范例证明,理论提供了数据挖掘的议题。

挖掘“产业领袖”就是在网络中找出一个重要的结点(node)。究竟是它的中心性(degree centrality)更重要,还是中介性(betweenness centrality)更重要?抑或是投资额大更重要?如果我们以扎根真相为目标,再寻找最接近目标的算法,将会使算法更为精确。

我们采用定性研究的德菲尔专家访谈法来收集扎根真相。研究团队先将所有的 VC 的 k-shell 值计算出来,^①其取值最高达 14,最低为 1,将其由高到低列成表,交由四位业界公认的专家(其中大多是“产业领袖”的执行长)去评比。最后得到 100% 认可的 42 家 VC 被选出来,定为扎根的“真相”。其中有 k-shell 值低至 7 的 VC 被勾选出来,也有一些 k-shell 值较高的却未入选,这说明如果用单一指标去挖掘“产业领袖”,误差将会很大,和真实世界中业界专家的认知会有很大的不同。

以这 42 家 VC 作为扎根真相,数据挖掘得到的算法纳入了五个网络指标,即度中心性、H 指数、k-shell 值、特征向量中心性、网络局域排序(local rank),以及两个非网络指标,即行业集中度和行业热度。用

① k-shell 是衡量网络中某个结点的重要性的指标之一,计算方法如下:将整个网络中与所有其他结点只有一条连结的结点剥除掉,那些被剥除的结点的 k-shell 值就是 1;然后在剩下的网络中又将与所有其他结点只有一条联结的结点剥除掉,那些被剥除的结点其 k-shell 值就是 2……以此类推,直到整个网络中所有结点都被剥除,每一个结点在第几轮剥除,其 k-shell 值就是几。

这7个指标做系统聚类分析,在42家“产业领袖”中找到35家,整网预测的准确率(accuracy rate)为0.965,召回率(recall rate)为0.83(Zhou, 2017),高于以任意指标无扎根真相下的挖掘结果。^①

当我们挖掘出产业领袖并对其进行了解后,回到上面的案例,比如,当风投网中的上千家公司分成了十群时,我们就可以诠释其结构,了解各社群的属性。比如哪几个群是某一领袖的圈子,哪几个群是两三个领袖抱团组成的圈子,哪几个群是好多个圈子聚群在一起成为拥有小世界网络结构(Watts, 1999)的较大社群,等等。^②从这些领袖的属性中我们也可以理解某个群到底是以外资VC为主,还是以国有资本或国内民营风投为主;弄清它们的投资对象是以单一产业为主,还是多元投资为主,以及包括哪些产业等问题,从而让我们对产业网的结构有进一步的认识。

这个范例说明理论提供了数据挖掘的研究议题,而在理论指引下进行的定性研究则提供了扎根真相,可用来校准数据挖掘的结果。

当然,圈子理论提供了一个解析产业网络结构的视角,但也并不排斥其他竞争性理论提出其他的视角,比如理性选择理论认为投资者会因资源互补而结合(Lockett & Wright, 2001),又比如社会网理论认为投资者会为了取得更好的网络结构位置而选择合作伙伴(Chiplin et al., 1997)。我们也不排除这些竞争性的理论可能会整合在一起,进而提出对产业网结构更有解释力的视角。

当我们可以用算法精确地找出产业领袖时,接下来的问题就是:为什么这些VC会联合投资?他们会找谁联合投资?这些是产业网结构中出现一个又一个集群的根本原因。化成更简单的问题就是,这些产业领袖会以什么标准来挑选他们在下一轮投资中的合作者。

四、数据挖掘与理论发展

若要回答在VC产业网中两两结合的联合投资关系因何而来的问

① 这样的算法仍有很大的改进空间,而且如果有更多次的专家意见,则可以提供更多的校准资料,使我们的算法更加精确。但是这也足以说明扎根真相在数据挖掘中的作用。

② 意即社群中包括多个相对独立的圈子,但在圈子间有密集频繁的联系,在中国VC产业中,经常就是这些产业领袖们扮演着圈间桥的角色。

题,数据挖掘可以发挥作用。在上述的资料库中,网络关系性指标就可以演绎出 81 个,而这些关系指标中哪些更为重要?研究者用 SBFG^① (structural balance based factor graph model) 模型计算出来的 7 个最佳预测因子分别是:相同国别、共同邻居数、中介中心性、关系距离、相同产权、投资领域数量以及相同的投资领域数量(Zhou et al., 2016)。分析这样的挖掘结果,我们可以看到有两类指标影响力特别大:一是网络结构指标,分别占据 2、3、4 位;二是同质性(similarity)指标,占据了 1、5、7 位。这项发现和对美国的生物科技风险投资企业的研究发现不太相同,后者同质性的重要性被其他变量控制了,使其结果并不显著(DiMaggio & Powell, 1982)。这说明中国风投产业的联合行为和美国不完全相同。第 6 位是投资领域数量,不难想象,一个多元化投资的 VC 会成为很多风投企业的联合对象,因为不同投资领域的投资者都可能向它伸出橄榄枝。

基于这些数据挖掘结果,嵌入理论(embeddedness theory)被引入进来,以解释为什么会有这样的联合行为。嵌入理论包括了关系性嵌入(relational embeddedness)以及结构性嵌入(structural embeddedness)(Burt & Burzynska, 2017)。葛拉堤(R. Gulati)在战略联盟研究中带入了关系性嵌入理论,指出两个战略伙伴间合作的次数越多,培育的信任就越强,合作的默契度也越高,所以下一次合作的概率也因之提高(Gulati, 1999)。由此可以得到如下假设:

假设 1:两个风险投资机构曾有过合作经验与再次合作的可能性正相关。

这一假设在中国风投产业的圈子现象中也有其合理性,一个圈子的核心 VC 需要一群特别紧密的伙伴,这是它的“团队”。在企业家创业过程中,往往机会出现时需要强有力的团队迅速行动,动员潜在的资源,把握机会(Granovetter, 1995)。同时,一个紧密群体在高不确定性环境中可以因内部的密网产生的监督作用降低道德风险,第三方的信任也可以作为可信赖的承诺使合作中的交易成本降低。所以,频繁的合作者会带来更频繁的合作,成为一个产业领袖圈子中的核心成员。

结构性嵌入则指的是一个行动者在社会网中的位置会影响其行

^① 该文比较了数个算法,结果 SBFG 在预测联合投资的精确度及收敛速度上都最优,其计算方法请参考 Zhou et al., 2016。

动与行动的结果。在探讨两个行动者的合作关系时,我们需要注意两个行动者间相对的结构位置带来的影响。关系距离会对两者的合作产生影响,一是因为信任可以传递(Burt & Knez, 1995),一步距离是直接有过合作经验者,知根知底自然更容易再次合作,而二步距离则是朋友的朋友,因为朋友的背书,第三方信任会使得二步距离也有一定的信任,较易合作。二是因为朋友常物以类聚,朋友的朋友相似性可能较高,在不同社交场合遇在一起,经中介变成直接关系的可能性也较大,所以合作的机会较多(Granovetter, 1973)。距离越远,信任的传递效果越差,成为直接朋友的可能性也越小,我们由此得到第二个假设。

假设 2:关系距离与未来联合投资关系形成呈反向关系。

共同朋友的数量会成为数据挖掘出来的重要因素(第二重要),因为信任传递效果迅速衰竭,三步以外就没有效力了,也就是朋友的朋友加以背书就不再具有可信度。同时,三步之外的人聚在一起直接认识的概率也直线下降,因此合作机会趋近于零。由于共同朋友数越多,信任传递效果越强,进而会面的机率也越大,因此合作的可能性也就更大。由此我们得到如下假设:

假设 3:两个 VC 间如果距离在三步以上,也就是没有共同社会网中的朋友,则合作的可能性大大降低。

将假设 2 与 3 放在中国 VC 产业网的圈子现象中,可以看到,在高度不确定性环境中要不断寻找更多的投资机会,所以圈子的核心除了需要有亲密的核心成员外,也需要大量的弱连带以接触更多的机会,并使关系圈时紧时疏,以带来不同的资源(Granovetter, 2002)。多一些合作过的伙伴,这些朋友的朋友就可能带来不同的资源,从而捕捉更多的投资机会。因此我们可以看到,圈子的核心会建立多圈层的关系网,既有亲密的核心成员,也会有较外围的圈内人,一方面掌握动员能力,一方面对更多的机会保持开放(Luo et al., 2017)。

研究者以 2000 - 2010 年两两 VC 间的合作频次、关系距离以及是否有共同联合投资伙伴作为自变量,控制了一些基于以前的理论设定的控制变量,如累积优势和投资领域相似度等,预测这两两 VC 间在 2011 - 2013 三年间合作的可能性,结果上述三项假设都成立(Luo et al., 2018; Luo & Zhou, 2014; 罗家德、曹立坤、郭戎, 2018)。

上述范例说明了数据挖掘的成果可以为理论的建构带来启发,而

理论建构的过程仍来自与其他理论的对话、逻辑的演绎、假设的提出以及资料的验证。

五、理论指导下的动态模型

简单地通过上述圈子理论与嵌入理论演绎出的因果模型去建构预测模型,可以依此一因果机制推论出谁和谁会建立合作关系,推论到不同时间点上和不同产业中,也可能推论到类似的文化环境内。但是大数据的引入会推动理论的重大发展,主要展现在研究复杂动态的系统,如果要进入这个研究领域,则需要建立网络演化的预测模型。

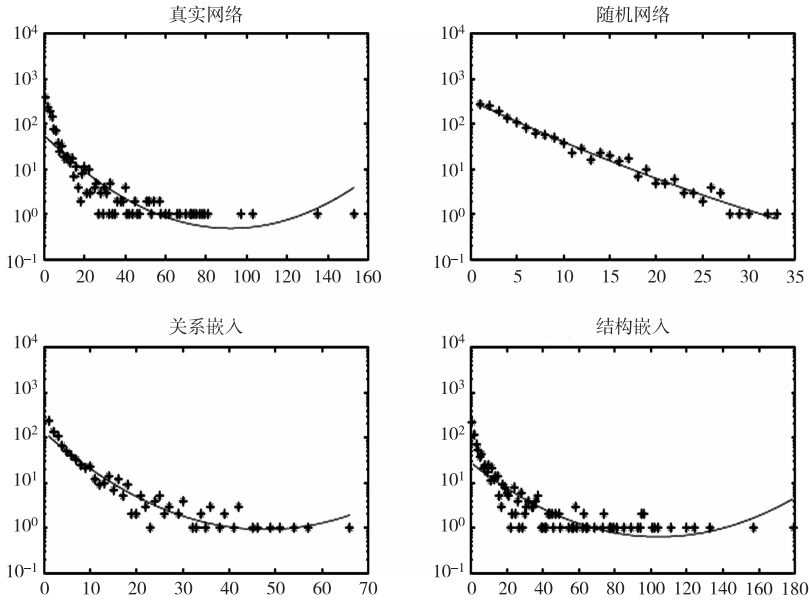
如果没有理论指导,网络动态模型往往会将一些基本的网络统计量,如网络规模、成长率以及网络密度等作为控制变量,再让网络内的结点随机生出与其他结点相连的线,形成基本的随机图(random graph)模型。然后将一些有趣的网络统计量作为自变量,比如三角闭合(closed triads)的数量及成长率等加入模型之中,看看新模型比基本的模型在预测未来网络结构上准确度会提高多少。

在风投产业中,因为有投资者与被投资者,所以研究者建了一个二模随机图的网络(2-mode random graph)(Gu et al., 2017),模拟真实的VC产业网络从2000-2013年共14年14期的投资。起始年2000年有75家投资者、375家被投资公司,依照真实网络算出来的统计量,两者都以每年30%的成长率增加规模。投资者进一步被分为三乘三的九类,在投资频率上分成高、中、低三类,在联合倾向上也分成高、中、低三类,依照真实网络统计量得出高投资频率者一年投资5.047次(模型用一期五次),中频者一年0.796次(也就是五期四次),低频者一年0.26次(四期一次)。有了这些控制变量,每一期投资者都会依其投资频率随机投向被投资者,当两个投资者同时投给一位被投资者时,就得到一次联合投资,将这些联合投资积累起来,就成为一张全产业的联合投资网,这是基本模型,是为模型一。

当圈子理论与嵌入理论被带进模型时,首先投资不再是随机的,而是主投者邀请跟随者加入联合投资,并依照假设1、2、3邀请特定的对象。另外,三类投资者按联合投资倾向分高、中、低三档,高联合倾向者每十次投资中会有九次邀人联合投资,中联合倾向者每五次投资中三

次有合作伙伴,低联合倾向者每五次只邀人合作一次。在每一期模拟中分两轮,第一轮投资者随机投资被投资者,第二轮主投者依照上述规则邀请其他投资者联合投资,在假设一的关系嵌入理论之下,过去合作的频次多则未来合作的机率也高,是为模型二。

模型三则把结构嵌入的假设 2 与假设 3 加入,三步距离以外者设定合作概率为零,两步距离者则低于有直接关系者,有直接关系者则依照假设 1 设定机率函数。研究者跑了 14 期的模拟模型,最后比较累积的真实网络,得到以下的结果:显然,在宏观网络指标中,如图 3 所示,所有 VC 的度分布模型二和模型三都明显地优于模型一,也就是关系嵌入模型与包含了关系嵌入及结构嵌入的模型都明显地比随机模型更接近真实网络。而模型三又比模型二的拟合程度更好。






注:转自 Gu et al., 2017。此图中每一子图的横轴是节点的度数,纵轴是该度数的节点数量。

图 3 宏观层面网络指标——度分布的比较

在网络微观层面的比较所有模体 (motif) 的统计数量,基本模型的预测能力都极差,而模型二与模型三则大幅度地改善了网络结构的预测能力,也就是关系嵌入模型与包含了关系嵌入及结构嵌入的模型预

测到的各类型模体数量都明显比随机模型的预测值更精准。而模型三又比模型二的表现更好(Gu et al. ,2017)。

微观层面网络指标的比较

模体形态	真实网络	随机网络	结构嵌入	关系嵌入
	7668	634	6646	7566
	633815	14642	441325	469410
	0	551	0	32126
	58765	127	51026	53060
	10258	50	5041	5359

资料来源:Gu et al. ,2017。

简言之,如果只是以网络统计量作为模型的预设值去预测网络动态演化,效果并不好,在这里理论的加入改变了预测准确度。上述范例显示,当圈子理论与嵌入理论加入后,模型二与模型三对 VC 产业网的演化预测比只有网络统计量的模型一精确得多。

回到复杂系统动态与非线性的演化研究上,这样的网络动态预测又有什么样的意义呢?

举例来说,曾有研究指出,银行互相拆借网络可以用来预测金融海啸,但是基于随机图得出的银行拆借网中,其几项模体(例如三角闭合)并无异常,仅在 2008 年金融海啸正式爆发时才不正常地增加。但加入了小世界网络理论(Zhou et al. , 2017)之后,在一个个“洞穴”(caves,意指被割裂的小团体)中边缘银行都是向核心银行拆借,而核心银行则是一个一个“洞穴”间的桥,一方面联结了分立的小团体,一方面又抱团成为一个精英小圈子,圈子内的核心银行紧密又频繁地拆借,核心银行再在自己的“洞穴”中拆借给边缘银行。在这样的模型中,三角闭合在三年前就显示出了异常(Squartini1, 2013)。

将银行拆借网的三角闭合数量作为预测金融危机的预测因子,可以吗?当然,这样的指标在方法论述中仅是资料挖掘的结果,它还需要发展成理论,而且它离解释复杂系统演化过程中的涌现现象还有相当长的路要走。一是这些指标只是在模拟模型中进行数据挖掘得来的,挖掘的结果并不具有推论的能力。二是对这些指标的诠释还需要发展

成理论,有待更多的理论对话来推动有效预测模型的建立。三是如金融风暴这样的涌现现象是网络结构与行为共同演化交相影响的结果,此一现象相关行动者的行为、行为的动机、行为的演化都付诸阙如,所以我们对此一系统非线性演化的因果机制尚缺乏理解。四是理论及因理论而来的预测模型还需要更多资料才能检验其有效性与可靠性,这只是一个孤例,需要通过更多类似的涌现现象资料的收集与验证,才能让预测模型具有可推论性。虽然在图2中“3”的过程里,从个体行为“集成”到宏观集体行为的研究长路才刚刚开始,但网络动态演化的预测模型却为这类研究踏出了坚实的一步。

六、总结:数据挖掘、理论与预测模型三者的互动

这篇论文的宗旨在于呈现计算社会科学的研究方法,如图1所示的数据挖掘、建构理论以及预测模型的三角闭环是其关键。无论是大数据还是结构化数据的数据挖掘,与社会科学理论和调查方法,以及与因果机制或系统动态模拟的预测模型三者之间对话才能完成一个具有推论性质的大数据研究。单单做数据挖掘只是整个研究闭环的一个起步,可以进行短期的、实用性的预测,但不足以建构理论,而通过理论的演绎才可以进行推论。

本文以一个VC产业的分析过程为范例,展示了图1的三角对话过程。“理论指导的扎根真相”部分展现了圈子理论如何为VC产业网的数据挖掘设定寻找重要结点算法的议题,同时在理论指导下以定性的方法去收集扎根真相,从而监督寻找算法的过程,并校准了算法的精确度。“数据挖掘与理论发展”部分又倒过来,展现数据挖掘对理论的贡献,数据可以作为验证理论的资料,同时通过对挖掘成果进行诠释,与其他理论对话,可以为理论建构提供启发,从而发展出新的理论。“理论指导下的动态模型”部分中探讨了如何用理论建立预测模型,并以预测模型还原真实资料。我们将假设1、2、3设定的因果机制纳入网络动态模拟模型,证明了圈子理论与嵌入理论能让VC产业网络演化预测模型的精确度大大提升。

当然,建构理论与预测模型的过程中仍有许多不完善的地方,比如寻找重要结点的算法还有很大的改进空间,只以现有的扎根真相而言,

就有7个产业领袖没有计算出来,因此,找出他们、找到理由、修改算法是下一步可以开展的工作。同样,引入圈子和嵌入理论的网络动态模型预测能力也有改进的空间,理论仍需继续发展,模型也仍需继续修正。

除了圈子理论与嵌入理论在产业网研究中继续发展外,其他竞争理论的对话也可能会启动更多对话,比如在“数据挖掘与理论发展”部分中对联合投资关系的原因的数据挖掘中,就发觉相似性在中国VC产业中十分重要,也应当发展出理论与预测模型,看看哪一个理论的预测能力更好。当然,也不排除将不同的竞争理论综合在一起其模型预测能力可能会更好,也许会发展出一个整合的理论。

还有一些发现的问题以及推论出来被验证的新“事实”值得深入探讨,往往新一轮研究会带来意想不到的理论修正。总之,以上形形色色的理由都使我们的研究又启动了新一轮如图1所示的三角对话,找到新的议题,收集更多的扎根真相,做更多的数据挖掘,针对新的挖掘成果进行诠释,开展理论对话,发展假设,验证理论,再用修正的理论指导新预测模型的建构,用新资料验证新预测……如此一轮又一轮,使理论不断改善,预测模型更加精确、推论范围不断扩大。

不过,我们还是要认识到,“大数据对理论发展的影响”部分所标示的理论新发展仍只是一个起步,还有漫漫研究之路要走,本文只是一个呈现研究方法的论文,虽然触及了网络动态模型的建立,但如欲解释并预测一个复杂系统的非线性演化,也就是诸如金融风暴、重大创新、社会运动、思想浪潮、制度变革、政经社会系统转型甚至革命等“涌现”现象,这样的研究是远远不够的。如上所述,我们还需要用理论去解释相关行动者的行为、行为的动机与演化,更需要研究网络结构和其动态演化与行为的交互作用,以及它们如何共同影响整体系统。虽然任重道远,但大数据的加入却使得相关资料的收集变得可能,从而开启了理论建构的新蓝海。

早期的大数据分析基于短期实用的目的,其研究策略是把收集到的数据当母体,以描述性统计和相关分析为主,或拟合一些既有的模型,这样的数据挖掘出来的预测因子及行为模式推论能力有限。我们只有要回答了为什么和如何的问题,才能作出更广泛、更精确的推论。

而现在计算社会科学把社会科学理论带入大数据分析之中,二者互补共进、相得益彰。一方面理论为大数据分析带来了更多的议题,理论指导下的定性、定量研究则提供了校准数据挖掘所需的扎根真相。另

一方面,大数据提供了验证理论的资料,同时数据挖掘结果可以启发新理论的建构。而被验证的理论则可以指导预测模型的建立,并能知道推论的边界在哪里,从而在边界之内用预测模型去预测更多的新“事实”。

在将来的研究中,大数据的加入可以拓展社会科学理论的新疆界,也即在图2中所示的“3”的过程——个人行为如何“集合”成集体行为、如何转化成宏观的场力,尤其是在一个非常态的复杂社会系统中如何产生涌现现象,最后带来系统的非线性转化。理论家们一直只能臆测,却因资料很少,难以对理论进行验证,从而很难更深入、更细致地发展与修正理论,更不用说提出预测模型预测这些现象。大数据的出现为这类理论拓展新疆界提供了可能性。

三角对话集合了计算机学者擅长的大数据挖掘以及社会科学学者擅长的理论与定性、定量研究,更需要加入能够建构复杂动态模型的人才。所以跨学科的整合,尤其是文理兼备、能够居中对话的人才成为大数据分析的关键,对研究社群最大的挑战将会是如何敞开学科门户的壁垒、敞开胸襟相互学习。

因为大量的人类行为发生在网上,网上行为和现实生活也有了可以推测的联系,而每天人们在网上留下的电子印迹都被记录着,大数据已经就在那里了,但如何正确地以计算社会科学的方法来善用大数据,是对我们提出的严峻挑战。

参考文献:

- 费孝通,1998,《乡土中国》,北京:北京大学出版社。
- 罗家德,2012,《关系与圈子——中国人工作场域中的圈子现象》,《管理学报》第2期。
- 罗家德、曹立坤、郭戎,2018,《嵌入性如何影响VC间的联合投资》,《江苏社会科学》第4期。
- 罗家德、樊瑛、郭晴、周建林、刘济帆、李睿琪,2018,《如何找出风险投资中的产业领袖——兼论计算社会科学中的扎根真相》,《探索与争鸣》第7期。
- 罗家德、秦朗、周伶,2014,《中国风险投资产业的圈子现象》,《管理学报》第4期。
- 罗家德、王竞、张佳音、谢朝霞,2008,《社会网研究的架构——以组织理论与管理研究为例》,《社会》第4期。
- 杨国枢,1993,《中国人的社会取向——社会互动的观点》,台北:桂冠图书公司。
- Burt, R. 1992, *Structural Holes: The Social Structure of Competition*. Cambridge: Harvard University Press.
- Burt, Ronald S. & Katarzyna Burzynska 2017, “Chinese Entrepreneurs, Social Networks, and Guanxi.” *Management and Organization Review* 13(3).
- Burt, Ronald S. & M. Knez 1995, “Kinds of Third-party Effects on Trust.” *Rationality and Society* 7.

- Chiplin, B., K. Robbie & M. Wright 1997, "The Syndication of Venture Capital Deals: Buy-outs and Buy-ins." *Frontiers of Entrepreneurship Research* 21(4).
- Coleman, James 1990, *Foundations of Social Theory*. Cambridge: The Belknap Press.
- DiMaggio, Paul J. & Walter W. Powell 1982, *The Iron Cage Revisited: Conformity and Diversity in Organizational Fields*. New Heaven: Yale University Press.
- Dunbar, R. I. M., V. Arnaboldi & M. Conti 2015, "The Structure of Online Social Networks Mirrors Those in the Offline World." *Social Networks* 43.
- Eagle, Nathan, Michael Macy & Rob Claxton 2010, "Network Diversity and Economic Development." *Science* 328(1029).
- Fu, Xiaoming, Jar-Der Luo & Margret Boos (eds.), 2017, *Interdisciplinary Social Network Analysis*. NY: Taylor & Francis Group.
- Galison, Peter 1987, *How Experiments End*. Chicago: University of Chicago Press.
- Granovetter, Mark 1973, "The Strength of Weak Tie." *American Journal of Sociology* 78.
- 1985, "Economic Action and Social Structure: The Problem of Embeddedness." *American Journal of Sociology* 91.
- 1995, "The Economic Sociology of Firms and Entrepreneurs." In Alejandro Portes (ed.), *The Economic Sociology of Immigration: Essays in Networks, Ethnicity and Entrepreneurship*. NY: Russell Sage Foundation.
- 2002, "A Theoretical Agenda for Economic Sociology." In R. C. Mauro, F. Guillen, P. England & M. Meyer (eds.), *The New Economic Sociology: Development in an Emerging Field*. New York: Russell Sage Foundation pp. 35 - 59.
- 2017, *Society and Economy—Framework and Principles*. Cambridge: Harvard University Press.
- Gulati, R. 1999, "Network Location and Learning: The Influence of Network Resources and Firm Capabilities on Alliance Formation." *Strategic Management Journal* 20(5).
- Gu, Weiwei, Jifan Liu & Jar-Der Luo 2017, "Analysis on the Dynamic Evolution Model of Joint Investment Network." Paper presented at 2018 International Network for Social Network Analysis Annual Conference, Beijing, May 30-June 4.
- Hempel, Carl G. 1966, *Philosophy of Natural Science*. NJ: Prentice Hall.
- Kosinski, M. W., L. H. Yilun & J. Leskovec 2016, "Mining Big Data to Extract Patterns and Predict Real-life Outcomes." *Psychological Methods* 21(4).
- Lakatos, Imre 1980, *The Methodology of Scientific Research Programmes: Volume 1: Philosophical Papers*. Cambridge: Cambridge University Press.
- Lin, Nan 2001, *Social Capital: A Theory of Social Structure and Action*. NY: Cambridge University Press.
- Lockett, A. & M. Wright 2001, "The Syndication of Venture Capital Investments." *Omega* 29.
- Luo, Jar-Der 2016, "Guanxi Circle Phenomenon in the Chinese Venture Capital Industry." In Jenn Hwan Wang (ed.), *Social Capital and Entrepreneurship in Greater China*. NY: Routledge.
- Luo, Jar-Der & Ling Zhou 2014, "Why Do Chinese Venture Capitals Invest Jointly—An Analysis of

- Complex Investment Network.” Paper presented at Academy of Management 2014 Annual meeting, Philadelphia, Aug 1 -5.
- Luo, Jar-Der, Ray-Chi Li, Fang-Da Fan & Jie Tang 2017, “Mining Data for Analyzing *Guanxi* Circle Formation in Chinese Venture Capitals’ Joint Investment.” In Xiaoming Fu, Jar-Der Luo & Margret Boos (eds.), *Interdisciplinary Social Network Analysis*. NY: Taylor & Francis Group.
- Luo, Jar-Der, Rong Ke, Kun-Hao Yang, Rong Guo & Ya-Qi Zou 2018, “Syndication through Social Embeddedness: A Comparison of Foreign, Private and State-owned Venture Capital (VC) Firms in China.” *Asia Pacific Journal of Management* (<http://doi.org/10.1007/S10490-017-9561-9>).
- Padgett, John F. & W. W. Powell 2012, *The Emergence of Organizations and Markets*. NJ: Princeton University Press.
- Prigogine, I. 1955, *Thermodynamics of Irreversible Process*. NY: Ryerson Press.
- Popper, Karl 1965, *The Logic of Scientific Discovery*. NY: Harper Torch Book.
- Powell, W. D. , D. R. White, K. W. Koput & J. Owen-Smith 2005, “Network Dynamics and Field Evolution: The Growth of Inter-organizational Collaboration in the Life Sciences.” *American Journal of Sociology* 110.
- Rubin, Donald B. 1974, “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66(5).
- Seager, W. 1995, “Ground Truth and Virtual Reality: Hacking vs. van Fraassen.” *Philosophy of Science* 62.
- Squartini, Tiziano, Iman van Lelyveld & Diego Garlaschelli 2013, “Early-warning Signals of Topological Collapse in Interbank Networks.” *Scientific Report* (DOI: 10.1038/srep03357).
- Viktor, Mayer-Schönberger & Cukier Kenneth 2014, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. London: John Murray Inc.
- Wasserman, S. & K. Faust 1994, *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
- Watts, Duncan 1999, “Dynamics and the Small-world Phenomenon.” *American Journal of Sociology* 105(2).
- Zhou, Yun, Zhiyuan Wang, Jie Tang & Jar-der Luo 2016, “The Prediction of Venture Capital Co-investment Based on Structural Balance Theory.” *Transactions on Knowledge and Data Engineering* 28(2).
- Zhou, Jianlin, Qing Guo, Ying Fan, Jar-der Luo & Weiwei Gu 2017, “Identifying the Leaders of Chinese Venture Capital Industry—Network Analysis VS. Ground Truth.” Paper presented at 2018 International Network for Social Network Analysis Annual Conference, Beijing, May 30 - June 4.

作者单位:清华大学社会学系
责任编辑:杨可