计算社会科学如何扎根真相①

——以如何找出中国风险投资的“产业领袖”为例

□罗家德[1]　樊　瑛[2]

内容摘要　计算社会科学与大数据的结合，将扎根真相指导数据挖掘，一方面有助于修补资料挖掘的不足，另一方面也可以将社会科学研究的议题与理论指导算法。以如何在中国风险投资产业中找到产业领袖为例，建基于在复杂网中寻找最具有影响力结点的方法，首先，需要基于业内访谈和德菲尔法，得到关于“产业领袖”的定义和业内公认的名单，用以验证之后数据挖掘的模型结果是否有解释力。然后，用清科数据库的VC共同投资数据建立动态的产业结构网络，计算各项网络指标。最后，利用数据挖掘的方法找到由动态数据直接进行分组的合理方案，从而验证有现实意义的寻找“产业领袖”的指标。在社会科学的理论指导下进行定性以及定量资料收集，提供了大数据资料挖掘的扎根真相，这使得过去做声音、图像、地景之类的大数据研究有了更多议题、方法和理论上的发展空间。计算社会科学的核心方法正是不断地在社会科学理论指导下的算法与社会科学方法上的扎根真相之间往往复复地对话，使得算法越来越接近扎根真相。

关 键 词　产业领袖　网络分析　扎根真相　风险投资　资料挖掘

作　　者　1罗家德，清华大学社会学系教授、博士生导师，清华大学社会网络研究中心主任。（北京　100084）2樊瑛，北京师范大学系统科学学院教授、通讯作者。（北京　100875）

基金项目　清华大学校内自主科研项目“基于通讯数据的关系强度与社会资本挖掘”（20175080105）；腾讯研究项目“以微信及QQ大数据分析个人人脉”（20162001703）

计算社会科学中的扎根真相①

计算社会科学（computational social science）在近年获得长足发展，利用新的计算方法和工具，系统地发展了社会科学一些概念的新衡量方法，比如社会地位、角色、团体形成过程中的社会关系，以及其中变动的身份认同等。[1]大数据的收集使得社会科学的研究方法产生极大变化，根据人们在网上留下的足迹可以归纳其网上行为的规律，推论其实际社会生活中发生的行为，进而构成调查资料、二手资料之外另一个庞大的资料来源。

然而，大数据过去以资料本身为母体，不强调随机抽样，不重视因果推论而只强调相关分析的资料挖掘方法[2]，固然有实务应用的价值，但只能挖掘出一些预测变量与行为模式，在方法论上以归纳法得出来的结论是不能作因果推论的，必须要对资料挖掘结果进行诠释，演绎发展出因果理论并进行理论验证，才能得到作因果推论的理论模型。[3]换言之，是理论建构了因果，资料挖掘可以告诉我们“是什么”，但无法回答“为什么”，从而无法作出有效推论模型，比如在挖掘什么是中国风险投资产业中（Venture Capital，简称VC）风投机构找寻联合投资伙伴的预测变量时[4]，为什么相同国别、共同邻居数、中介中心性、关系距离、相同产权类型、相同投资领域数量等是与联合投资意愿相关性最高的变量？为什么像

共同邻居数是如此重要？在何种情境下它会变得不重要（理论模型的边界在哪里）？它如何影响联合投资意愿（影响机制）？它们之间会不会有中介变量或共变关系？

进一步，理论又是在相互检验中发展出科学社群所能认可的因果规律。拉克托区[5]的科学方法论提出了研究纲要建构与竞争的概念，要求建构真正能在研究纲要中当作硬核（hard core）的理论。基于硬核，我们可以发展出针对一特定现象的解释模型，并用同一社群认可的可衡量的“事实”在“一定显著水准下证实或证伪”解释模型，这是一个理论硬核的保护带，“证实或证伪”只是为了理论间“交互验证”，以说服更多理论的追随者。硬核的追随者大多是不会改变的，因为保护带中模型指定可以修正，方法可以更改，甚至还可以换个模型指定，最后当一个理论模型换了几个指定都找不到显著的结果时，这个现象就不解释了，而换个现象又能发展新的理论模型。但无论如何，都不会动到硬核，只是当一个硬核的追随者越来越少时，理论间的竞争就出现了胜败。换言之，是理论建构与理论间的对话、竞争方才建立了因果机制，进而可以作因果推论，大数据资料挖掘只是建立理论过程的一个起步，挖掘结果并不足以建构因果机制与推论模型。

计算社会科学与大数据的结合，一方面，修补了资料挖掘的不足，使挖掘结果更趋近于相关科学社群所认可的“事实”；另一方面，也可以将社会科学研究的议题与理论指导网上数据挖掘与建模，从而以验证一些社会科学的理论假设，比如一个地区的社会网多元化与经济发展间的关系[6]，也可以从计算社会科学的算法中作出推论性统计，从而产生预测的效果。比如，从个人的电讯资料中可以挖掘出个人的人脉圈层，也就是不同亲疏远近的个人中心网[7]。同样地，在得到电信数据的演算法后，可以推及其他人的电信往来，得到他们强弱连带的不同圈层。 晚近，这些计算社会科学的发展改进了过去大数据研究以数据挖掘[8]中的描述性统计与相关性分析为主的研究，改以社会科学的理论，如社会地位、角色、社会关系、身份认同、大五人格、邓巴圈等，设定数据分析的议题，指导数据中的各类指标的计算以及建模过程，从而得到了与理论相呼应的算法。但是在这个过程中却出现了一个重大的问题，就是算法和理论指涉的现实现象到底是否稳合，这就是扎根真相（ground truth），以及用扎根真相修正算法的过程（ground truthing）。[9]

扎根真相简单地说，就是相关科学社群以共同认可的研究方法收集到的“事实”，其概念来自于遥感(remote sensing) [10]，指涉卫星与高空照像借助雷达或X光等手段，最后成像与地面的真实事物到底有多大差距。这概念引入了数据挖掘与计算社会科学领域后谈的就是，数据中的各类指标以及这些指标组合成算法最后与理论指涉的现实概念到底有多大落差？修正算法的过程就是一个不断逼近现实的过程。仍以电信数据计算邓巴圈为例，电信数据中有的只是一个人打了多少人的电话，接了多少人的电话，每一个通话对象的通话频率、通话时长、间隔分布、通话地点与时间等指标，要如何组合才能算出两人间的连带强度？邓巴圈就是由强到弱分成五圈，得到扎根真相最好的方法就是去调研分析对象，让他们自己说哪些连带强，哪些弱。进而，强弱连带的指标包括了互动频率、关系久暂、互惠内容以及亲密程度[11]，可以通过问卷调查扎根真相。换言之，社会科学理论不仅设定议题，指导算法，也是社会科学的实证方法，不管是定性访谈还是定量调查，都是进一步验证算法效度的手段。

如下将扎根真相指导数据挖掘，找到算出“产业领袖”的方法，并以中国VC产业中联合投资网络资料加以验证。

如何扎根真相

风险投资自在中国发展以来，就逐渐显现出与西方不同的投资趋势与投资逻辑。相比较来说，中国VC更倾向于通过建立关系和圈子来完成VC的联合投资。一方面，对于整个中国VC界来说，都存在着一种“主投-跟投”的投资模式[12]，主投者占据着绝对的资源和市场信息优势，从而拟定投资计划并主导投资，而跟投者往往由于在整个VC关系圈子的外围而无法得到有利的信息，只能作为跟随者“分一杯羹”。基于这样的主投-跟投模式，国内的风险投资产业中存在“圈子现象”：一些实力相对较弱的投资

者会围绕在较为强大的投资者周围，形成以大型投资者为核心、以联合投资为纽带的圈子。圈子现象表现出几个特点：其一，它反映了中国人以自己为中心建构的差序格局人脉网的现象，也就是因亲疏远近不同而有圈内人、圈外人之分，圈内人又有核心与边缘之分，不同亲近程度的合作伙伴会受到不同的待遇。其二，虽然每一个VC都在建立自己的圈子，但只有那些有实力的、在大多数投资中都作主投者建立的圈子才具有影响力，成为众多VC都想争取合作、加入圈子的对象。另外，虽然重要的投资者都有自己的圈子，但整个风险投资产业却是一个小世界网络，也就是说在圈子之间往往会有“桥”将之联系起来，而扮演“桥”的角色者往往是重要圈子的领袖。圈子现象表明，在中国的VC大环境下，政策不确定性、信息不对称造成的高风险等问题，使得中国的新兴VC机构更倾向于通过建立关系来获得行业声望和地位，从而长期获得较好的地位和利益回报。一些老的、大的或有特殊资源禀赋的风投机构建立的圈子，成为新兴或较小VC竞相加入的对象，这些能建立自己庞大圈子的风投机构崛起为产业领袖。

就操作性定义而言，产业领袖就是一家风投公司在它的绝大数投资中都是主投者。因为每一笔投资中，往往只有一个主投者拟定投资计划，并代表投资人进入被投公司的决定机制中。罗永胜、李远勤[16]利用2000—2014年的风险投资数据，来研究风险投资网络核心-边缘结构的动态演进规律，发现参与风险投资的机构形成一个网络，其中大多数VC处于网络边缘，核心区成员个数稳定且占据主导地位；但该研究仅从理论和数据上进行了核心和边缘的区分，并没有针对基础事实上的检验。圈子研究进一步发现，这样的大型投资者中许多已成长为“主投者”，带领跟投者进行联合投资[14]并在合作中起主导作用，“主投-跟投”的合作模式有助于新兴的VC机构（绝大多数是跟投者）长期受益。基于社会科学理论和业内共识，笔者希望找到能够有效度量“产业领袖”，即具有产业重大影响力的主投者的社会网络指标。

（一）从定性访谈中定义产业领袖

作为一种重要的行为策略，联合投资越来越被国内外的风险投资机构普遍使用。一般来说，联合投资行为分广义上的联合与狭义上的联合：广义上的联合指两个或以上的机构共同对某一特定项目进行了投资，不同的机构可以在不同时间介入该项目；狭义上的联合投资指两个或以上的机构在同一轮融资中对某一特定项目进行了投资。[15]在对业内人士的访谈中，我们专门询问G对于“产业领袖”的定义。他并没有给出一个确定的定义，而是提了几条建议：“千万不要用核心这个词，在社会网络中找一个词比较合适。风险投资领域在业界里称呼很少，容易得罪人。”我们对产业领袖的定义主要来源于对VC行业的基本了解，即在大多数投资中都是主投者，并且在声望、资源占有和影响力上对业界有重大作用的VC机构。例如Y在访谈中提到，“股东的研究机构强，政策资源多……一般都是他们领投并且排他”。因为有主投-跟投现象，并且经常只有一个董事会席位，所以只有一个主投可以占这个席位。主投写投标书，可以占有有利资源，然后寻找合适的跟投，从而达到共担风险或是共同收益的目的。其中，共担风险往往意味着跟投者要放弃短期收益，而共同收益的收益分配也有很大的不平衡。

（二）根据Delphi法寻找产业领袖

一是专家意见调研。首先，我们利用滚雪球的访谈方法，访谈了多位业界专业人士，并在他们的推荐下特别挑选对全产业有较深了解的人士，结束前邀请他们用口述的形式填写一份名单。名单上列明了2000—2013年中出现过的所有VC名，为了方便专家的勾选，并将其按照k-shell值由高到低的顺序排列（该顺序由于后来的数据清洗方法变动可能会有所不同，但这并不影响被访谈者的勾选）。访谈员先解释了前述定义的“主投者”，当受访者被问到“是否认同该VC在业界中具有较高影响力”这一问题，根据受访者口述对名单的看法，访谈员在旁边记下该受访者对每个VC的看法（这里被提到的VC大部分截止在k-shell=6，受访者就会停止名单浏览）。需要说明的是，我们虽然假设k-shell值作为产业领袖的测量指标，但在访谈时尽可能多地让受访者勾选了VC，表达了他们对其是否为产业领袖的看法，从而减小名单对k-shell值指标的直接影响。主要的看法分为三种：认同、还不错、不认同，分别将其记为Y、G和N三种。此外，受访者分别对他们认为的一些具体VC

机构的形式和性质进行了补充，我们在后续名单分析中将其标准化。值得一提的是，这里的名单提供时将本土VC和外资VC完全分开，因此在勾选时能够有较大的公正性。

二是定性资料处理。首先，我们对拿到的名单进行了基本的处理：(1)名单中去掉PE、天使投资和券商直投。(2)原有数据中Y表示“肯定其大佬地位”，N表示“否定其大佬地位”，空白表示被略过不发表意见，G代表Good(地位还不错)。(3)对除了填字母以外的补充说明(如“是并购”、“近几年地位不好”等词汇直接默认为肯定其地位)；填“母基金”由于机构在数据库中体现的数据都是和VC之间的联合投资，因此将其默认为是VC并且肯定其为大佬。(4)将外资VC中k-shell值为14(排名最高)的所有VC的空白处都填上Y(因为这样的VC“江湖地位”很高，有的受访者会表示同意而不再在问卷上填写)。具体在筛选时为了统一标准，我们将方法进行如下整理：(1)将G视作Y，都表示肯定其为大佬。(2)而对不清楚、不熟这一类的否定补充，将其视为放弃评判，也就是空白。(3)将在填空处填上PE的机构，因为他们在我们的数据库中是只计算作为VC的联合投资网络的，所以可以忽略不计(只要已有的机构类型是VC就可以)。(4)对于填写“现在不活跃”的VC，因为我们想要说明的是一个动态的数据，那么这个VC曾经是有名的，我们也将其视作“江湖地位”的一种表现，即“Y”。(5)只要在填写处出现一个N，就不把它放在“大哥”名单中。(在本土里，不存在一个人填N剩下人都填Y的情况，顶多是一N两Y一空白，所以这样的处理可以保证公平)。经过将原始的勾选名单标准化后，可以得到一个统一的被分为三级的“产业领袖”名单。

表1 三级“产业领袖”的筛选标准

	本土	外资
一级	至少75%都是Y，无N	100%为Y
二级	50%是Y，无N	50%为Y，无N
三级	只有一个Y/G，无N	无

(三)扎根真相

虽然在名单中显示出了三级，但为了研究的准确性，我们只选取一级领袖为准确的领袖名单，并将二级和三级视作名单勾选中的干扰项予以排除。在此基础上，我们得到关于中国VC业界“产业领袖”的名单。其中，有29家外资和13家本土VC，利用这样的名单，我们将清科数据库中清洗后的数据的全部VC分为“领袖VC”和“非领袖VC”两组，用以验证之后的资料挖掘结果是否有解释力。

如何验证算法

(一)数据收集与处理

有了扎根真相作比对标准，研究进入资料挖掘过程，以寻找一系列指标与一套算法去计算出谁是产业领袖。一方面，我们以清科数据库2000—2013年的数据为基础，为联合投资网的生成需要同时在网上搜索资料加以补充②；另一方面，为了保证投资过程中联合投资动机的准确性，将VC投资轮次分为天使(种子)期、发展期、扩张期和成熟期，去掉天使(种子)期和成熟期部分的数据，以及和PE联合投资的情况，共收集到8426条数据记录，从而得到了具有联合投资经验的673家VC机构，并且建立他们的整体合作网络。为了描述风险投资公司在风险投资网络中的重要地位，我们使用五种具有代表性的无向网络中节点重要性排序的指标：

一是度中心性。“中心性”是社会网络分析中的重要概念之一，Bavelas曾对中心性作出开创性的研究，认为行动者在网络中的位置越接近核心，其影响力越大。度中心性是最直观也是最常用的概念：“度(degree)”的概念来源于图形理论，描述在网络图中一个节点与其他多少个点有连带(tie)。[16]我们采用狭义的联合投资概念，即当且仅当两家风险投资机构在同一轮风投中对一家企业进行投资，才称两家机构间有过一次“联合投资”。那么这种联合投资网络可以用邻接矩阵A来描述。邻接矩阵$A=\left(a_{ij}\right)_{N\times N}$是一个N阶矩阵，矩阵元素取值为0或1。$a_{ij}=1$当且仅当节点$v_i$与节点$v_j$之间存在一条连边时，表示两家投资机构$i$与$j$之间存在联合投资关系；$a_{ij}=0$，表示两家投资机构之间不存在联合投资关系。在风险投资网络中，度描述了一家机构曾经与多少其他机构有过联合投资行为。度越高的投资机构就有过越多的联合投资行为，在整体行业中拥有更多的合作者。节点i的度是指在网络中与节点i直接相连的节

点的数目，具体可定义为：$k_i=\sum_j a_{ij}$。在风险投资网络中，一家风险投资公司的度可理解为与该风险投资公司拥有合作关系的风险投资公司数目。度值越大，说明该风险投资公司的合作者数目越多，在一定程度上也能说明该公司的实力更加雄厚。

二是k-shell值。k-shell分解[17]是一种基于节点度的粗粒化排序方法，可以用于判定节点在网络中所处的位置。首先，网络中度中心性为1的节点可视为最不重要的节点，当去掉这些度为1的节点后，网络中就会出现一些新的度中心性为1的节点，再将这些度为1的节点及其相连的边去掉，直至网络中没有度中心性为1的节点为止。此时，所有被去除的节点及其它们之间的连边，被称k-shell值为1，而剥去了k-shell值为1之后的新网络中，每个节点的度中心性值至少为2。以此类推，直至网络的所有节点都被去除，此时网络中每一个节点都被划分到相应的k-shell中，这样就得到整个网络的k-shell分解。节点的k-shell值越高，说明节点越处于网络的核心位置，节点的影响力越大。在联合投资网络中，K-shell值较高的机构意味着位于联合投资网络中较为核心的位置，对整体网络拥有较强的影响力。

三是H指数。节点i的H指数[18]可定义为一个最大值h，满足节点i至少拥有h个度值不小于h的邻居。H指数是一个网络局部指标，不仅考虑了自身的邻居数，而且还考虑了邻居的能力。具体到联合投资网络中，这意味着VC本身并没有很大号召力，但由于其合作伙伴的地位很高，它的地位也随之上升。这一点在行业知识上来说，能够成为一个有效的检验标准。

四是LocalRank。LocalRank是一种基于网络局域信息的排序算法，是一种对度中心性的扩展，考虑了节点的四阶邻居的信息，节点i的LocalRank值可定义为：$\begin{aligned} LR(i)&=\sum_{j\in\Gamma_i}Q(j)\\ Q(j)&=\sum_{k\in\Gamma_j}R(k)\end{aligned}$。其中，$\Gamma_i$表示节点$i$的一阶近邻，$R(k)$表示为节点$k$的一阶近邻及二阶近邻数。

五是特征向量中心性。特征向量中心性是一种基于网络全局信息的排序算法，其基本思想为：一个节点的重要性不仅取决于其邻居节点的数量，而且也取决于其邻居节点的重要性。特征向量中心性指标是网络邻接矩阵对应的最大特征值的特征向量。记x_i为节点i的重要性度量值，则$x_i=c\sum_j a_{ij}x_j$，其中c为一比例常数，$i=1,2,\cdots\cdots,N$。上式可表示成以下矩阵形式：$\vec{x}=cA\vec{x}$，$\vec{x}$为矩阵A的特征值c^{-1}所对应的特征向量。

除了网络指标之外，我们还选取了与产业相关的产业集中度和行业热度两个指标。前者指该VC机构在2000—2013年间的所有联合投资中，共涉及多少家不同的行业；后者指该VC机构在2000−2013年间的所有联合投资中，所投行业是否在热门行业中。至于与投资绩效最相关的IPO率、投资数、投资总额等数据，由于数据缺失尤其是在产业领袖上的缺失非常多，再加上绩效本身就是很大的决定性因素，因此没有作为分组指标。

（二）数据挖掘与分析

基于清洗好的2000—2013年的联合投资网数据，我们构建了含有999个节点、2265条边的无权无向的风险投资网络。该网络中的节点代表各个风险投资公司，节点之间的连边表示两家风险投资公司之间存在联合投资关系。该网络的最大连通子图含有601个节点以及2217条连边，通过访谈获取的中国VC界“产业领袖”均包含在最大连通子图中。为了方便指标度量以及使得识别算法更加高效，我们仅关注风险投资网络的最大连通子图并计算其包含的节点的指标值。

根据以上的网络指标，我们希望能够找到一种分类方法，使得它的分类结果在统计上与Delphi法得到的名单最为接近，即最为有效。在评价分类效果的时候，一方面使用计算机领域中常用的三个指标来进行评估：（1）准确率计算的是被正确分类的样本数与总样本数之比。（2）精确率计算的是所有“正确被检索的item”占所有“实际被检索到的”的比例。（3）召回率计算的是所有“正确被检索的item”占所有“应该检索到的item”的比例。另一方面，我们利用Ⅰ型错误和Ⅱ型错误两个指标来讨论被误判的VC。其中，Ⅰ型错误又称第一类错误，指拒绝了实际上成立的假设，为“去真”的错误；Ⅱ型错误称第二类错误，指不拒绝实际上错误的假设，为“存伪”的错误。Ⅰ型错误意味着有VC在名单中，但没有被分组结果放进目标组；Ⅱ型错误意味着有VC本不在名单中，但

被分组结果放进了目标组。

我们首先基于三个网络指标：度、H指数、k-shell值对601家风险投资公司做聚类系统分析[19]，并将所有风险投资机构（VC机构）分为三组。通过将其与“产业领袖名单”进行比对可以发现，当仅利用网络指标对其分类时，仍有21+7家产业领袖被排除在外，同时有2家不应是产业领袖的VC机构。也就是说，这样的分类标准并不够准确。进一步分析产业领袖的属性我们发现，一些VC只在少数领域中较为活跃，或是只投资最热门的行业，从而虽然其在圈内的联合活动并不活跃，却能够为大众所认可。因此，根据这一部分产业领袖的行业特征，我们加入了行业集中度和行业热度。需要注意的是，这里的热门行业选取的是2000—2013年所有VC机构投资的不同行业的次数中，被投资次数最多的top10行业。通过计算，我们得到的热门行业分别为：B2C、应用软件、其他无线互联网服务、新能源、其他机械制造、环保、网络社区、新材料、网络游戏、IC设计。与此同时，我们考虑了两个不同的网络指标——特征向量中心性、Local Rank，以及两个非网络指标——行业集中度、行业热度。我们根据这7个指标对601家风险投资公司重新做系统聚类分析，结果发现这次的计算结果明显好于之前的分类结果。第三组明显找到了9家产业领袖VC，而第一组是明显的“目标非领袖组”，在第二组和第三组合并的情况下，我们可以近似地得到完整的“目标领袖”分组。最后，我们可以计算评价分类效果的准确率、精确率和召回率。

通过对上述分组的判别和有效度检验，可以发现这样的分组最接近我们的要求，被判断出的VC的总体情况也最符合我们之前对产业领袖VC的定义。

（三）去真与去伪

我们已经看到了数据挖掘在找出VC界产业领袖过程中体现出的优势，同时也得到了较有说服力的结果。然而，依然存在14家非领袖进入“目标领袖组”，同时有7家领袖VC被分入了“目标非领袖组”。虽然我们尝试加入其他的指标，但拟合效果并没有得到显著改善。因此，我们选择进一步分析被误判的VC机构本身的特质和数据特性，从而简要解释其被误判并且难以改善的原因。

7家被分入了“目标非领袖组”的VC，根据成长历史及现有资料的分析又可以分为两类：一类以外资企业为主，这一类企业本身的投资专业度非常高，由于在某一重要产业领域中地位极高而被业界公认为产业领袖。但是，由于其极高的专业性和极窄的投资范围，在网络指标和产业指标上的表现均不出色。也就是说，产业领袖VC在圈内拥有极高的地位，但它在寻找伙伴时几乎不会广泛进行筛选，拥有较为封闭的联合投资圈层。因此，在没有和投资金额及IPO率等绩效数字直接相关的资料判读下，我们很难通过机器学习来找出这一类产业领袖。同时，因为数据库本身的信息限制，该类外资企业的投资金额、IPO率等数字缺失较多，因此无法将其加入到指标之中。另一类以本土VC为主，由于其庞大的资源（往往是政府资源）和很好的投资表现而被业界认可，而在后期由于政府政策的调整发生了角色转换，由直接参与投资的VC变为了对被投公司间接产生投资影响的PE机构。与此同时，还可能成立一些子公司来参与直接投资，这使得我们在数据统计时会发生偏误，也会使其在圈层中的地位逐渐减弱，取而代之是名下一些子投资公司的崛起。

与上述“去真”情况相对的14家非领袖VC被误判进入了领袖组可以分为三类：第一类是一些老牌的所谓外资，其掌权者和经理人团队都来自于国内，只是在上世纪90年代初期、国内VC刚开始兴起时回国投资，以其较为先进的视角领先于国内新兴VC，因此在前期占有很大优势，从而体现出较好的网络指标数据和产业数据；但后期由于真正外资开始不断涌入国内，同时国内一些VC也开始崛起，这部分VC的势力逐渐衰弱，以至于到现在业界已经不承认其“产业领袖”的地位。第二类是一些大型国有企业的子公司，在国有企业的资源和资金支持、授权下进行定点定向的投资，其联合投资行为既没有市场逻辑，也没有圈层逻辑，纯粹依附于其母公司的控管，类似于“券商直投”的企业，只是在登记信息时被记为VC机构，可以在研究中忽略。第三类是纯粹误入的非领袖VC。这类VC以T公司为代表，利用人脉网络和圈层与业界领袖形成多次联合投资，但因其自身实力并不强大，也并没有主导投资和控制资源的能力，因此并不被业界认可为业界领袖；但也因其强大的人脉关系和圈层关系网络，其网络指标和所涉及行业的广

泛程度，给人以“产业领袖”的假象。

这几项去真与存伪的误差中最常见的案例就是，曾经的产业领袖正在没落中，或是原来的跟随者正在冉冉升起，被认可为有潜力的产业领袖，这些需要进一步以网络动态变化的指标加以解释，是我们在产业网中寻找产业领袖的下一步研究。而计算社会科学的核心方法正是不断地在社会科学理论指导下的算法与社会科学调查方法得到的扎根真相之间往往复复地对话，使得算法越来越接近扎根真相。

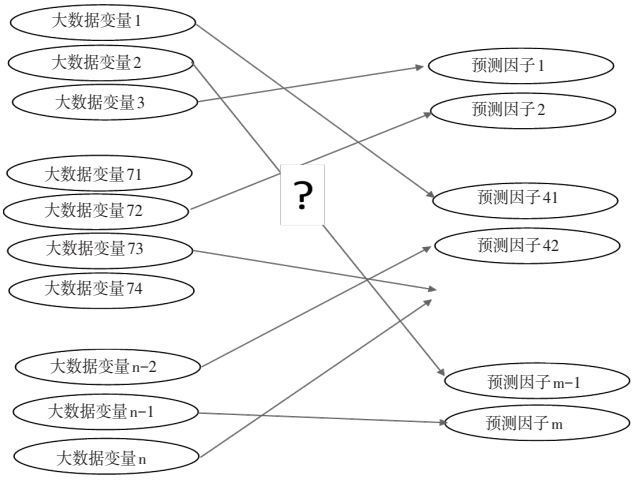
大数据研究的新发展：算法扎根真相的议题、方法与理论

在社会科学的理论指导下进行定性以及定量资料收集，提供了大数据资料挖掘的扎根真相，这使得过去做声音、图像、地景之类的大数据研究有了更多议题、方法和理论上的新发展。

（一）理论驱动的大数据研究的议题空间

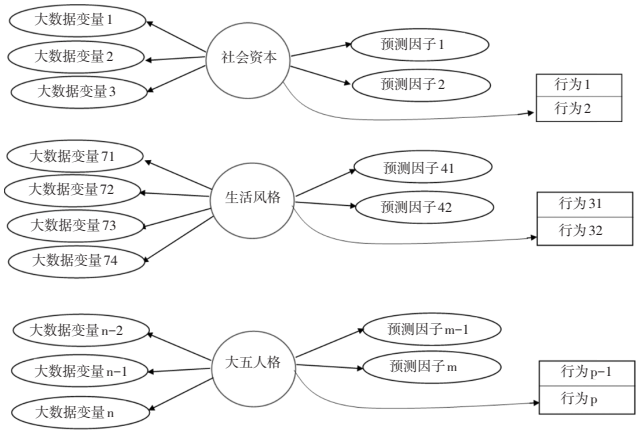
早期的大数据研究聚焦在应用上，做描述性统计和相关性统计期间，多维度交叉整合来分析不同的预测因子。如图1所示，把收集到的数据当母体，不强调随机抽样；主要做描述性统计和相关分析，不强调因果推论。这样以数据挖掘为主的大数据研究方法，往往回答了“是何”（what），而不能回答“如何”（How），即了解过程发展的机制以及“为何”（why），或者说无法得到因果关系，找不到这些大数据变量和预测因子之间的内在关系。

图一 data-driven的大数据分析



新一阶段的是理论驱动的大数据研究，采取了扎根真相的大数据挖掘技术，用社会科学理论指导大数据精准分析的议题，也用社科理论指导分析方向和建模方法，进而以大数据方法刻画用户画像，比如个人的人脉、社经地位、社会资本、移动规律、大五人格与生活风格等个人图像。换言之，社会科学理论的功能之一是为大数据研究提供了新议题。如图2所示，这些理论又能使大数据研究进一步聚焦在应用价值上，预测出个人图像，再通过个人图像研究人的行为与态度，像行为1、行为2等可以是商业类的，如借贷行为、采购行为、逛网行为，网上交友行为、信用能力等，从而可以作精准营销、精准推送的研究；也可以是犯罪行为、社会支持行为、互助行为、心理疾病等，从而对心理健康、社会组织培育、人口迁徙、社会信任问题等一系列社会治理问题进行研究。

图二 Theory-driven（社会科学理论）的大数据分析



社会科学的研究方法，定性的或定量的调查数据或二手数据正是扎根真相的主要来源。这是社会科学理论驱动的又一个功能，可以判断哪些变量跟社会资本有关，哪些变量跟生活风格有关（不同社会科学理论可当作中间变量来构成桥梁，产生不同预测因子），以此得出较为准确的行为预测分析结论。

（二）计算社会科学中扎根真相的方法空间

计算社会科学在计算机学、社会科学和系统科学之间建起合作的桥梁，对于传统计算机学科中的数据挖掘的描述性统计提供了事实基础和理论依据，提高了其分析结果的可信度和现实性。过去的系统科学界和计算机学界缺少社会科学的理论基础和讨论方法，即使计算出了结果也很难寻找有力的现实解释。而社会科学尤其是社会网络学派，拥有统计方法和理论基础，但面对大数据的新时代背景和动态网的研究趋势，却没有行之有效的数据分析和模

型建构方法。计算社会科学的研究方法将二者结合，可以展示跨学科合作的进展和成果。就计算社会科学来说，扎根真相的概念在方法论上有着重大的意义。

第一，它提供了更加有效的方法来分析由大数据生成的结构化数据，使得社会科学的知识有了和大数据进行对话的切入点。简单地说，是社会科学的理论提供了我们大数据挖掘的议题，比如研究挖掘出“产业领袖”的计算方法，就是基于风险投资行业中总有很多圈子，而圈子是一种社会网结构，即围绕在一个顶点周围因亲疏远近而产生数个圈层的结构。顶点就是“产业领袖”，所以分析VC产业网络结构时一个重要的工作就是寻找这些顶点。

第二，除了设定议题之外，社会科学的定性或定量调查方法提供大数据资料挖掘比对的扎根真相。比如，在德菲尔法的访谈中，我们让业界中一些熟知业内生态的人选出“产业领袖”。在方法论上，单纯的资料挖掘不足以建构因果机制，从而很难作出推论，扎根真相提供了相关学术社群共同认可的“事实”，同时其背后往往也提供了相关的理论基础。比对“事实”，资料挖掘才意有所指，更为精确；而有了理论，才说明了因果机制，便于作出因果推论。

第三，在社会科学的理论指导下，与产业领袖最可能相关的指标被计算出来，并被放入资料挖掘的模型中。当然，在这一轮的数据挖掘中我们得到的算法精确度仍有改善的空间。所以在下一轮的研究中，基于这次数据挖掘结果，被误判为领袖的14家VC被拿来分析，以了解误判的理由，找到可能的指标；接下来我们需要收集更多的扎根真相，验证新的算法。如此一轮又一轮的理论指导→扎根真相→数据挖掘→得到算法并加以验证→与理论对话得到指导……直到得到让我们满意的资料挖掘的成果，从而推论到不同的时期或不同的产业中的资料。

概言之，计算社会科学结合了社会科学的理论与研究方法以及大数据资料挖掘的方法，一方面纳入大数据分析作为验证与发展理论的重要工具；另一方面指导了大数据的资料挖掘，使之有的放矢，扎根真相并校正其算法，从而以理论建构出分析结果背后的因果机制。

（三）中国经济领域的社会网络理论空间

在一个有领导者–跟随者的环境中，在中国往往围绕着领导者会形成一个关系圈子[20]，风险投资产业因为每一笔投资都有主投–跟投之分，所以也会形成领导者–跟随者的圈子结构，领导者的个体联合投资网也会有差序格局现象[21]，使用算法找出领导者是描述这样的产业网的第一步。

继之，领导者身边的圈层也可以复杂网络的ERGM模型得出，一个跟随者是否入圈？是否成为班底？圈子现象有初识者、熟人与班底之分，但如何通过联合次数推断出VC之间的关系却并非易事。通常的方法可能是研究者根据经验确定一个阈值，但这无疑有很强的随意性。应用复杂网络研究领域中的ERGM方法，得出控制条件下的随机模型网络，并与真实网络进行对比可以发现，一次合作只是试错过程，合作两次以上才算作熟人或者说入圈，7次以上才属于班底/铁哥们儿；而且这一发现也在访谈当中得到了定性资料的支持。[22]

一方面，未来的研究可以在产业领袖与圈层的算法之下，划出一个个主要投资者的圈子，以及跟随者和这些主要投资圈子间的关系，从个人中心网的演化中去预测VC的种种表现，如营利能力、生存时长、所投公司的上市比率、兼并比率等。一个产业的整体网也是由这些圈子组成的，从不同圈子间的合纵连横、此消彼长则可以看到产业网络的演化，产业作为一个系统，其动态网演化是预测产业发展至关重要的一环。[23]

另一方面，对于VC界来说，利用动态网的发展机制，能够在未来预测哪一部分VC有实力和潜力成为产业领袖，这样的产业讯号对业界来说也具有很大的意义。基于中国VC产业界的联合投资倾向、“圈子”形态和“主投–跟投”的联合投资模式，我们能够更加精准地了解VC产业界的未来走向和绩效潜力。此外，将动态网络模型加入到社会网络的研究中，也为中国在经济领域的社会网络研究提供了一个新颖且有效的方法。

可以预见，如果拥有更多的数据支持，我们就能够给更多的大数据背后的社会科学问题提供更加切实可信的证据和解释。

注释：

① 感谢清华大学社会学系硕士郭晴、刘济帆，北师大系统科学学院博士生周建林、李睿琪共同参与此一研究与论文写作。

② 清科数据本身就是网上非结构化数据经过整理而来，为了分析联合投资网络的需要，研究团队另外从网上下载或比对了四千条数据。大数据这里指涉的是非结构化的电子印迹数据，往往体量庞大，在社会科学理论指导下，大多有一个将之整理为结构化数据的过程，才能为社会科学的分析所用。

参考文献：

[1] Evans, J. A. ,Aceves P. Machine Translation: Mining Text for Social Theory. *Molecular Microbiology*, 2016: 539−547.

[2] Viktor, Mayer- Schönberger, Kenneth, Cukier. Big *Data: A Revolution That Will Transform How We Live, Work, and Think*. London: John Murray Inc, 2014.

[3] Popper, K. R. The *Logic of Scientific Discovery*. London: Routledge, 1965.

[4] Zhou, Yun, Wang, Zhiyuan, Tang, Jie, Luo, Jar-der. The Prediction of Venture Capital Co- Investment Based on Structural Balance Theory. *Transactions on Knowledge and Data Engineering*, 2016: 537−550.

[5] Lakatos, I. *The Methodology of Scientific Research Programmes: Volume 1: Philosophical Papers*. Cambridge: Cambridge University Press, 1980.

[6] Eagle, N., Macy, M., Claxton, R. Network Diversity and Economic Development.*Science*,2010:1029−1031.

[7] Dunbar, R. I. M., Arnaboldi, V., Conti, M., Passarella, l. The structure of online social networks mirrors those in the offline world. *Social Networks*, 2015: 39−47.

[8] Walker, Saint, J. Big Data: A Revolution that Will Transform How We Live, Work, and Think. *American Journal of Epidemiology*, 2013: 181−183.

[9] Fu, Xiaoming, Luo, Jar-Der., *Margarete, B. Social Network Analysis: Interdisciplinary Approaches and Case Studies*. NY: CRC Press, 2017.

[10] Seager W. Ground Truth and Virtual Reality: Hacking vs. van Fraassen. *Philosophy of Science*, 1995: 459−478.

[11] Granovetter M. S. The Strength of Weak Ties. *Social Networks*, 1973: 1360−1380; Marsden P. V., Campbell K. E. Measuring Tie Strength. *SocialForces*, 1984: 482−501.

[12] 罗家德、邹亚琦、郭戎. 中国风险资本联合投资策略的差异化现象. 现代财经−天津财经大学学报，2016(4).

[13] 罗永胜、李远勤. 我国风险投资网络核心—边缘结构的动态演进. 财会月刊，2017(2).

[14] Luo, Jar-Der L. Guanxi Circle Phenomenon in the Chinese Venture Capital Industry. *Social Capital and entrepreneurship in Greater China*. NY: Routledge, 2016: 56−71.

[15] Brander, J. A., Amit, R., Antweiler, W. Venture−Capital Syndication: Improved Venture Selection vs. The Value - Added Hypothesis. *Journal of Economics & Management Strategy*, 2002: 423−452.

[16] Daley D. J., Gani J. M. *Epidemic modelling: an introduction*. Cambridge: Cambridge University Press, 1999.

[17] 张金柱. 利用 k-shell 分析合著网络中的作者传播影响力. 现代图书情报技术，2012：65−69.

[18] Lu, Linyuan, Zhou, Tao, Zhang, Q. M., Stanley, H. E. The H- index of a network node and its relation to degree and coreness. *Nature Communications*, 2016: 10168.

[19] Rokach, L., Oded, M., Clustering methods. *Data mining and knowledge discovery handbook*. NY: Springer, 2005; Johnson S. C. Hierarchical clustering schemes. *Psychometrika*, 1967: 241− 254; Kaufman L., Rousseeuw P. J. *Finding groups in data: an introduction to cluster analysis*. NY: John Wiley & Sons, 2009.

[20] Luo, Jar- Der, Cheng, MengYu, Zhang, Tian, Guanxi circle and organizational citizenship behavior: Context of a Chinese workplace. *Asia Pacific Journal of Management*, 2016: 649−671.

[21] 费孝通. 乡土中国. 上海：上海人民出版社，2013.

[22] Luo, Jar- Der etc. Mining Data for Analyzing Guanxi Circle Formation in Chinese Venture Capitals' Joint Investment. *Interdisciplinary Social Network Analysis*. NY: Taylor & Francis Group, 2017: 177−196.

[23] Powell, W. D etc. Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences. *American Journal of Sociology*, 2005:1132− 1205; Padgett J. F., Powell W. W. The Problem of Emergence. *The Emergence of Organizations and Markets*. Princeton: Princet on University Press, 2012.

编辑 李 梅