微博信息传播的量化研究（讨论版）

文 / 醒客工场微博研究组

本文是醒客工场研究员们关于微博传播研究的基本框架，并不是最终的正式版本，仅介绍基本研究思路、方法，供内部交流。醒客工场的项目研究，采用的是“迭代式”研究方式，所以这个“讨论版”，一方面是向大家汇报我们的研究进度，另一方面，是希望各位老师、朋友能给我们一些反馈意见和建议，以便我们进一步调整、完善——你们的热心支持，是我们前进的动力，谢谢。

1 研究背景介绍

1.1 整体研究框架综述

从微博问世以来短短 2 年多的时间，已在互联网和传媒界引起广泛关注。截止 2011 年底，微博用户已达 2.5 亿[1]，超过社交网站的用户数量。有研究表明，微博既有社交网络的特性，也具备一些传播媒体的特性[1]。对中文语境和中国信息环境下微博的使用的研究，目前还处于起步阶段，缺乏系统、科学、量化的研究。本研究项目作为更大规模系统研究的一部分，希望能起到抛砖引玉、见微知著的目的。

微博问世以来，产生了各种有意思的事件。无论是草根明星的出现，还是明星微博受到的关注，以及微博舆论与主流媒体在信息传播上的异同，都反映出微博的传播对社会经济政治文化生活的影响。一方面，各类认证用户，特别是娱乐、体育、政治、商业等领域名人通过使用微博，不仅完成了他们已有的社会权利向网络的转移，同时使得他们在原本专长领域（如娱乐或文艺）的影响力向其它领域（如公共事务）发生了扩张。另一方面，草根博主的出现，给了普通人在网络舆论空间的话语权，这是在 Web2.0 技术之前从未有过的。微博作为新媒体，参与了很多重要事件的报道，与主流媒体形成互补，为推进社会信息流动，起到积极的作用。

微博的信息，以简短、实时、广播式的特点，通过以单方“关注”机制形成的社交网络进行传播。微博作为新媒体的特性，既符合传播的普遍规律，又与以往的传播媒介有很大区别。在微博的使用中，用户可以感受到微博的传播有哪些特点？微博信息寿命多长？微博的响应程度与哪些因素有关？微博用户网络有什么特点？引爆点在哪里？这些都是我们重点关注的问题。

根据这些问题，醒客工场作为一家专注于信息化和互联网推动社会演进研究的独立研究机构，将从如下两个方向开展研究：

- 我们采用基于数据统计的实证研究方法。对数据进行统计，建立模型，用数据对模型进行验证，对模型原型进行逐步迭代，直到模型具有一定的普遍性。通过这种方法，来发现微博传播现象背后的规则。
- 我们采用跨学科的综合研究方式。来自信息科学、心理学、新闻传播学、数学等专业的醒客工场研究员们，将从各自的领域视角展开对于微博传播现象的分析，努力向各位读者呈现一个综合、立体的问题分析视角。

1.2 研究数据采集情况简介

我们通过新浪提供的 API 接口，分别随机抽取 1 万、5 万、10 万个源微博以及其所有转发记录、评论记录以及作者信息。我们发现，由于访问权限、数据传输稳定性

1 数字来自《中国互联网络发展状况统计报告》，中国互联网络信息中心（CNNIC）

等方面的限制，所抓取的转发、评论数据存在一定程度的缺失。具体如下：

表 1：转发数据缺失状况

	应有转发数	实有转发数	缺失率
1 万	15949	4436	0.722
5 万	175085	13852	0.921
10 万	185509	27601	0.851

表 2：评论数缺失状况

	应有评论数	实有评论数	缺失率
1 万	14174	12846	0.094
5 万	173310	64665	0.627
10 万	174054	122021	0.299

1.3 参与报告的研究员分工

本文其它部分组织如下：第二、三章介绍国内外微博研究的现状，第四章介绍本研究的理论基础，第五章展示具体应用领域的部分研究结果，第六章进行总结并展开讨论。参与报告撰写的醒客工场研究员具体分工情况如下：

表 3：参与本课题的研究员分工情况

内容主题	负责人
1 研究背景介绍	张鹏翼、常政
2 国外微博研究状况综述	张鹏翼
3 国内微博研究状况综述	虞鑫
4 微博传播的基础理论模型	
4.1 微博传播的半衰期研究	李岩
4.2 微博引爆点计算的数学模型	常政
4.3 微博传播的心理学模型	戴必兵
5 微博信息的应用统计分析	
5.1 新浪微博数据的相关基础分析	张鹏翼
5.2 微博的社会网络分析	虞鑫
6 研究结论和下一步规划	常政、张鹏翼
课题数据分析支持	袁晓燕
课题数据抓取	张雪峰
课题总协调	常政

2 国外微博研究状况综述

国外对微博的研究，主要以 Twitter 为样本，研究的内容包括微博平台作为新媒体的特性、其在一些特定领域（诸如应急反应和灾后救助）的新应用、以及微博平台作为社交网络对信息传播的影响和推动等。

2.1 微博作为社会化媒体

微博作为社会化媒体，其显著特点之一就是对新闻事件的报道和关注。这种报道和关注与传统媒体相比，有其自身的特点。西方有学者 [2] 将新闻的一个话题、观点、口号或用语等称为模因（meme），作为文化信息传承的单位。模因经过复制和模仿、变异和选择的过程在信息传播的过程中发生演化。对某个话题或事件的关注和追踪，可通过对模因的追踪来进行。

Kwak 等人 [1] 使用爬虫程序，抓回截止到 2009 年 7 月的全部 Twitter 微博，共抓得 4170 万用户、14.7 亿关系、4262 个流行话题和 1 亿 600 万条微博（tweets），再用程序去除 spam。他们关于 Twitter 整体状况的详细研究分析表明，Twitter 不仅是一个社交网络，更具备了新闻媒体的一些特征。例如：

- 热门话题的时效性很强，话题更迭较快。
- 少数用户可以直接达到大批受众（名人或媒体效应）。
- 大部分用户可以通过口口相传达到大批受众。

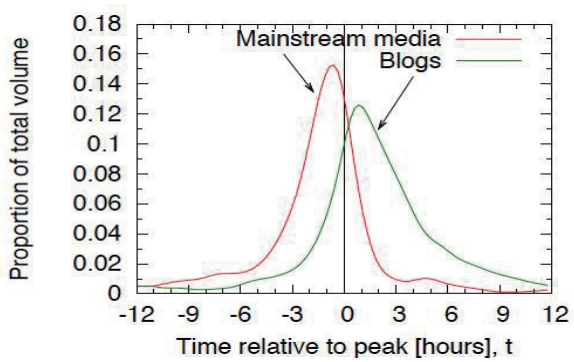

图 1：西方传统主流媒体与博客社会媒体的时间差 [2]

2.1.1 社会媒体与传统媒体比较

Leskovec[2] 通过对模因的追踪分析表明，博客作为一种社会化媒体，对事件报道的峰值通常比西方传统主流媒体晚 2.5 小时左右。传统媒体对事件的报道量或关注度在达到峰值前缓慢增长，之后则迅速降低；而社会媒体增长相对较慢但持续性较高（图 2）。此研究的对象主要是

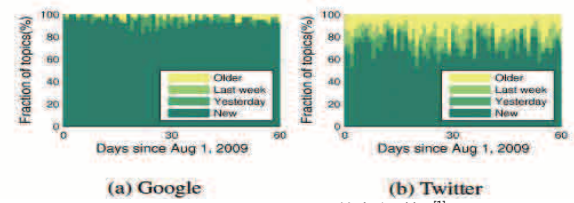

图 2：Google Trends 和 Twitter 的内容时长 [1]

博客，但其方法对微博的研究具有很大的参考作用。

图 2 中显示的 Twitter 与 Google Trend 内容的新旧情况的比较也与上述结论相吻合，即社会媒体对事件的持

续度较高，传统网络媒体则实效性更强。

2.1.2 社会媒体的滥用

由于微博作为社会化媒体，吸引了大量的用户群。而传统媒体也会对一些微博事件进行跟踪报到，使得通过微博这一平台，可以很容易就达到相当大规模的受众群，受众的教育文化社会背景各不相同。网络社会媒体的出现，令草根个人和草根组织更容易获得话语权的同时，也使得他们对社会媒体的滥用更加容易。少量通过集中控制的账户发出或制造关于某些话题的一些言论，混淆视听，甚至是制造虚假新闻。针对这种假装是自发、流行的草根行为，但实际上是某些组织或个人有策划预谋的行为，[3] 提出了监测和跟踪社会媒体中的政治滥用的方法和技术。

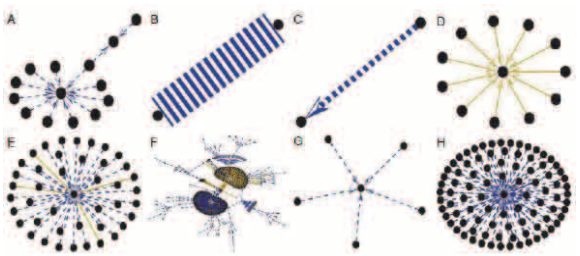

图 3 : 8 个不同话题的传播网络 [3]

图 3 介绍了其中的一个方法，通过描述话题在网络中的传播的特性来甄别滥用。上图上半部 4 个话题是滥用的网络形式，下半部 4 的话题是正常的话题传播轨迹。这就引出了微博作为社交网络对传播不可避免的影响，在 2.2 节中有详细介绍。

2.2 微博作为社交网络

Twitter 作为一个通过关注关系将用户联系起来的网络，具有社交网络的很多特征 [1, 4]；但与直观感受不同的是，研究表明“Twitter is not so social”，因为关注并不是一种相互关系，只有 22.1% 的用户互相关注，相比之下，其它社交网站的相互关注度为：Flickr 68%，Yahoo! 360 84%，Cyworld guestbook 77%，因此与其他网络的朋友（friends）关系不一样，Twitter 的社交网络不反映网下的人际关系，而更接近 RSS 订阅。同时这种单向的关注关系非常脆弱 [5, 6]，9 个月的时间中，节点平均失去了 39%的关注关系。

对微博的内容分析 [7] 也表明，微博的内容具有一定社会化特性，其中信息共享占 56%，会话占 37%，信息寻求 8%。分享个人信息和有方向性的对话占了一定比重。与此同时，微博作为社会媒体和社交网络的特性是相辅相成的，社会媒体区别于传统媒体的最大特征就是其网络结构。微博的媒体作用是通过微博作者的社交网络散布出去的，因此西方学者都非常重视研究微博的网络结构对信息传播的影响。

2.2.1 微博作者的影响力

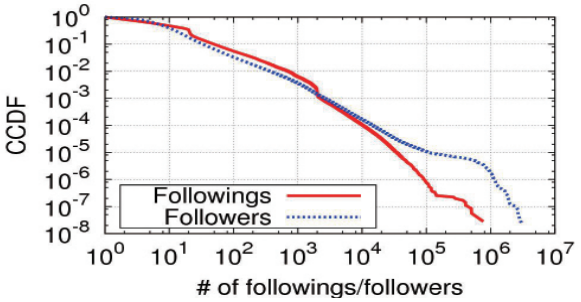

图 4 ：Tweeter 微博作者的关注和被关注数 [1]

在图 4 中，X 轴表示关注和粉丝的数目，Y 轴的互补累计分布函数 (CCDF) 则表示对连续函数，所有大于 a 的值，其出现概率的和，即 F(a)=P(x>a)。图 4 的红线表示关注的数量，它有两个明显的下滑出现在 x = 20，2000。20 是 twitter 推荐给新用户的用户数。而 2000 是 2009 年之前用户可关注的人数上限。蓝线表示被关注的数量，蓝线在 $x<10^5$ 之前体现的是一个幂指数分布，这跟一般的网络分布是一样的。而在 $x>10^5$ 之后，出现了一个凸起，这说明拥有粉丝数多余 10 万人的用户，拥有比幂指数分布预测的粉丝数目多得多的粉丝。这在一定意义上反映了社会权利向网络权利转移的过程中的马太效应。

2.2.2 转发的影响力

图 5 表明，小于 1000 个粉丝的时候，新增受众与初始响应范围无关。这表明转发机制给每个用户提供了相对均等的传播信息的机会。而大于 1000 个粉丝的时候，转发有可能带来更多或较少的新增受众。研究同时显示，50%的转发在 1 个小时之内，75%的转发在 1 天之内。对于一条微博的前第 6、7 个转发，每个转发者与前一个转发者之间的转发时间几乎是相等的，也就是说，微博转播的速度非常得快，有一点信息接力的意味。

综上所述，国外研究者对微博的研究主要以 Twitter 为研究对象，关注其作为社交网络和社会化媒体的双重特性。微博作为一个新兴的技术产物，以对人们获取和分享信息的方式产生了重大影响。传统人际网络的分隔度为 6，即每两个人可以通过平均 6 个人相连，电子邮件、

即时通讯工具等新技术形成的网络分隔度也大约为 6，而 Twitter、Facebook 等新兴 Web 2.0 网络分隔度都不到 5。新技术大大缩短了人们之间的距离，对信息的传播也必将带来重大影响。国外研究的方法和结论是否适用于中文语境和中国信息环境，还需要进一步的深入研究。

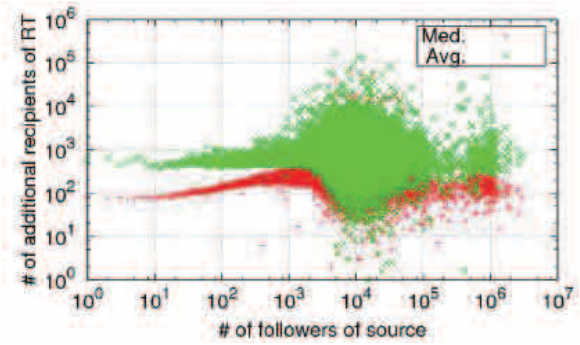

图 5：转发的附增受众 [1]
x 轴：源微博的粉丝（初始响应范围），y 轴：转 发带来的附增受众

3 国内微博研究状况综述

国内对于微博的研究主要集中在三个领域：信息传播机制、用户行为特征以及微博作为一种新媒介对其他事物产生的影响。

3.1 信息传播机制

微博信息传播的互动模式可以划分为“链状”、“环状”和“树状”。其中链状结构是指源微博和评论、转发形成热点话题；环状结构是指评论、转发的内容使得源微博向相关的话题扩展，从而加强了该话题的深度和广度；树状结构是指源微博发布后，评论与转发可能与源微博的话题毫不相关，用户采取了更加主动性的回应模式，同时树状结构也可能包括链状和环状的互动结构 [8]。

微博影响力的动力机制内生于基于平台所激发的内容协同生产以及基于用户间社会关系网络所打通的信息通路，达到了话语释放和群体联通的作用，微博影响力的本质是对信息资源的凝聚和整合 [9]。微博的“嵌套性”特征就是实现不同圈子节点之间互通互联的原因所在 [10]。

对于“节点”的概念，张佰明认为具有双重含义。首先，节点是微博中发布和接受信息的用户；其次，节点也是参与信息互动的用户及其呈现给其他用户的相关信息的结合体 [10]。这说明，在相同的信息环境下，不同的微博节点将呈现不同的作用和效果。

信息传播的面积、传播链长、传播的速度及传播的生命周期取决于事件本身、微博客空间内外的干扰因素、传播的节点以及节点的合理布局几个方面。此外，这项研究也验证了强势节点与信息传播的影响力具有正相关性，即强势节点出现越早，信息传播影响力越大 [11]。

一项针对网络话语的内容分析显示，网络言论更多代表的是中间阶层的“民意”，而居于社会下层群体的“产业工人”、“农业劳动者”则丧失了网络话语权 [12]。

微博在灾难性事件发生时，具有使得信息传递迅速、呈现全面、聚合专家等正效应，也具有冗余信息过多、信息失真、群体极化的负效应 [13]。

高承实等人对于微博舆情的监测体系进行了探索性的研究，他们将微博信息空间模拟为三维模型，分别由网民的参与信息（编码维）、舆情信息（抽象维）、传播媒介（扩散维）构成 [14]。

3.2 用户行为特征

微博用户的使用行为方面，源微博的转发数和评论数之间具有中度相关性，关注数、粉丝数、微博数三者之间均具有高度正相关性，一定程度上符合了线下人际交往的特性——积极主动的使用行为将会产生更多的响应 [15]。

有研究者观察发现，官员微博具有开通渠道、协商制度、表达观点、深化讨论和社会动员的作用，同时也存在选择性倾听和选择性回应、自我展示做秀的问题 [16]。

用户使用微博的心理机制来源于现代社会人们倾诉和寂寞的需求，微博“点对面”的传播特性，使得即使没有人响应，也可以满足使用者的倾诉 [17]。

3.3 作为新媒介产生的影响

社交型传播技术将会产生三种新的“新闻形态”，分别是私语式新闻、对话式新闻和直白式新闻，这些都是对传统新闻形态的解构 [18]。也有研究者认为微博的出现改变了传统媒体的文体、文风和表达方式，其本身——而非其所承载的内容——就已经改变了传播生态 [19]。

蔡骐提出了“粉丝型受众”这一概念，粉丝群体通过新传播技术建构了一个个打破“时空区隔”的，并且颇具亚文化特征的阐释性社区 [20]。

信息技术的发展逐渐颠覆了传统“真实”的概念：在网络中越是想让大脑获取无限信息，身体就越是被禁锢在极为有限的现实空间中。“虚拟真实”正在挑战传统的

现实真实[21]。

微博打破了传统媒体议程设置的单一局面，不仅带来了多元的话题信息，同时也使得新闻评论的价值取向趋于多元，反映了社会不同阶层和文化背景下的意见表达[22]。

4 微博传播的理论模型

4.1 微博信息传播的半衰期研究

本节主要对微博信息传播的半衰期、传播结束以及传播周期进行了科学的定义；并根据上述定义，通过数据统计求得了目前新浪微博中微博信息传播的平均历时半衰期。为对微博信息传播进行时间维度的研究建立基础，为微博信息传播的分析和应用提供初步参考数据。

4.1.1 微博信息传播时间维度研究的意义

微博信息传播是在移动互联网技术发展过程中出现的，以社会化网络结构为基础，适合通过手机等移动信息工具[1]进行实时发布、快速传播，融合了文本、图片、音视频文件、地理数据和社交 ID 等信息的，一种全新的信息传播方式。由于微博本身的技术基因，这种新的信息传播方式在内容形态、信息容量、消息结构、传播路径、传播速度、受众范围以及信息传播规律上，不仅大大不同于传统的信息传播形态；与桌面互联网信息传播相比，也出现了非常大的差异，它具有自己鲜明而独特的传播特征和规律。同时，微博信息传播与以往传统媒体的中心化广播式传播和传统互联网的去中心化对等网状传播都不同，其传播呈离散的多中心分级网络扩散传播形态[11]。由于微博网络的构建是基于社会化网络结构的，因此微博信息传播中综合具有人际传播、组织传播和大众传播三种社会传播结构，并呈现出与现实社会中不同利益群体话语权力结构的拟合。因此，在微博信息传播具有鲜明的独特规律的同时，它也成为了可以被数字化追踪的社会传播结构的标本。对微博信息传播的研究可以部分地展示出现实社会中信息传播的特点和规律。

在对微博信息传播的现象观察、过程分析以及规律总结中，我们看待这个信息传播“过程”，不可避免地将其视为一个与时间向量相关的运动现象。而从社会宏观角度看，微博信息传播中的各种话题的涌现和更替，更是一个社会热点信息的不可逆的新陈代谢过程。因此，从时间维度上对微博信息传播进行研究，具有多种学术和实践意义：首先，时间维度上的研究对于客观地分析社会传播中热点话题的老化更替，发现微博信息传播的动态规律有着基础性的作用。尤其是以定量分析方法来确定什么是热点话题，科学地定义热点话题的传播起点、传播结束和传播周期是研究微博信息传播过程的基石。其次，通过对热点话题的传播过程进行历时的观察，结合传播过程中对重要时间节点上的其它相关因素的分析，我们可以发现微博中热点话题的形成因素，从而部分地揭示形成信息传播热点的动力机制。再次，通过大量数据的统计分析，可以对微博信息传播的一般时间规律进行总结。而对任何一个微博信息传播过程中某一时刻的数据分析，将其与一般性规律比较，可以发现信息老化更替的状态。通过这种共时的观察，可以对其传播阶段进行预判。因此，从时间维度上科学地观察微博信息传播现象并进行量化定义，将非常有助于我们揭开这一全新信息传播形态的面纱，并部分窥测社会传播的客观规律。

4.1.2 微博信息传播的基本概念

以信息论的观点，微博信息传播作为一种客观的社会现象，其过程是可以被观察、测量和计算的。从研究角度，我们对于这一社会现象的各个环节和过程进行了重新定义，以排除由于主观认识上的区别，不同观测体系带来的测量、计算误差，尽量客观科学地研究其基本规律。

定义 1 微博信息传播：在某一个确定微博系统中（如新浪微博、或腾讯微博分别为一个确定的微博系统），以公开的方式进行的信息传播过程，为微博信息传播的研究对象。以非公开（如私信）或公开传播于其它媒介（如其它媒体中的微博精选等）均不属此项研究的对象。

定义 2 源微博：博主不通过微博系统的“转发”和“评论”功能，而直接发布的微博即为源微博。源微博的特点是，对其传播不会必然引起对其它微博信息的传播，即在传播过程中，可被回溯到源头的微博。

源微博与源信息之间的区别在于视角不同，前者是从传播过程的角度，是传播的起点；后者关注的是微博信息的内容特征，同一源信息可以产生多个源微博，而同一源

1 瑞士信贷 2011 年 5 月发布的《新浪微博调查报告》中统计，通过手机使用新浪微博的用户占用户总量的 34%，笔记本电脑用户占比 38%。2011 年 2 月未来资产发布的《新浪微博研究报告》中指出，2010 年四季度，新浪微博来自移动互联网的用户数为 40%。

微博也可以传播多个源信息。

定义 3 微博信息传播的起始：源微博的发布就是该微博传播过程的起始，通常用精确到秒的时间来表示。

定义 4 微博信息传播的方式

（1）强化节点传播，是指在微博传播中使用了系统的“@”功能，而实现对特定节点的强化，称为微博强化节点传播方式。

（2）泛节点传播，指在微博传播中没有使用系统的“@”功能，所有节点平等，称为微博泛节点传播方式。

（3）等值传播，指在微博再传播中，对原有信息内容不进行改变的传播方式。

（4）变值传播，指在微博再传播中，对原有信息进行修改、评论等，造成内容改变的传播方式。

（5）等值传播与变值传播均以信息形式作为判断标准，与信息量（信息熵）的改变没有必然联系。再传播与源微博发布相对应，包括转发、评论和对评论的转发和回复。

定义 5 微博信息传播的范围

（1）微博传播的初始范围，指源微博博主的粉丝用户与源微博内容中用@功能强化传播的用户的并集。

（2）微博传播的最大范围，根据研究范围和微博系统的星形网络结构，一条微博的传播最大范围是指，在该微博系统中自微博传播起始点之后，存在的用户集合的总集。

（3）微博传播的响应范围，指对源微博 X 进行评论、转发和评论转发的节点的并集，可以用不重复的节点数量来测量，计作 U_X。响应范围的数值意义非常类似于阿弗拉米斯库(A.Avramescu)方程[23]中的传播范围，即 C_0 值的正相关值 C_0'。

（4）微博传播的初始响应范围，指微博 X 传播的初始范围用户集与响应范围的节点的交集，可以用不重复的节点数量来测量，计作 U_0 。这不同于阿弗拉米斯库（A.Avramescu）方程的初始增量值 m。

（5）微博传播的递延响应范围，指微博 X 传播的响应范围中非初始响应范围的用户集，可以用不重复的节点为数量来测量，其数值为 $U_X - U_0$。

定义 6 微博信息传播的原点与节点

（1）源微博的博主在源微博发布时为微博信息传播的原点。

（2）对源微博通过评论、转发或对评论进行转发，来进行再传播的用户，包括再次通过评论、转发或回复等行为进行再传播的源微博的博主，都称为微博信息传播节点。

4.1.3 微博信息传播半衰期和传播结束的科学定义

为科学地从时间维度来比较和分析微博信息传播的特征和规律，我们需要在上述基本概念的基础上，为这一信息传播过程找到一个通用的基本指标。我们引入微博信息传播半衰期的概念，通过这一客观可观测可计量的指标，作为微博信息传播时间维度研究的基石。并以此数量统计定义的结果，来定量化定义微博信息传播结束的概念，进而可以求得微博信息传播周期的时长。

4.1.3.1 微博信息传播的半衰期

（1）微博信息传播半衰期的概念

定义 7 微博信息传播半衰期：微博信息传播过程中，自源微博发布至观测时刻止，最近更新的对源微博的再传播数量达到观测时微博传播的响应范围节点数的一半时，则从这一时刻到观测时刻的时间长度为微博信息传播的共时半衰期，简称微博共时半衰期；从源微博发布到这一时刻的时间长度为微博信息传播的历时半衰期，简称微博历时半衰期。

共时半衰期其数值表现了微博传播的趋势，数值越小则说明其传播趋势越强，信息老化越快，传播还在发展中。历时半衰期的数值表现了微博信息传播的初始速度，其数据越大说明传播起步很慢，反之非常小的数值说明微博信息在初期是迅速传播。

（2）微博信息传播半衰期的数学定义

设源微博 X 自 t_0 时发布，在 t_1 秒内有 c_1 个节点响应，进行了再传播。至观测时刻 t_n 秒内有 c_n 个节点响应进行了再传播。则 X 源微博在观测时的响应总节点数量为 $C_X=\sum_{i=1}^{n}c_i$，当某一时刻使 T 满足 $\frac{1}{2}C_X=\sum_{i=1}^{T}c_i$ 时，T 为微博 X 的历时半衰期，T_n- T 为微博 X 的共时半衰期。

为直观地理解微博半衰期的含义，区分历时半衰期与共时半衰期的定义，展示以及观测时刻与两种半衰期的关系。我们在 5 万条随机抽取的微博中找到响应总个数最多一条微博，统计出该微博每个小时的响应节点数。然后根据每个小时的节点数画出微博响应的趋势图，并标明历时半衰期、共时半衰期以及观测时间进行示意。（见图 6）

关于微博半衰期，这里有以下三点需要特别说明：第

一，这里所定义的微博共时半衰期，与观测者选取的观测时间密切相关。在微博信息传播过程的不同阶段，选取时刻不同，求得的数值不同。因此，反向也可以通过在某时刻求得的数值来判断，在这一时刻微博信息传播处在具体什么阶段。

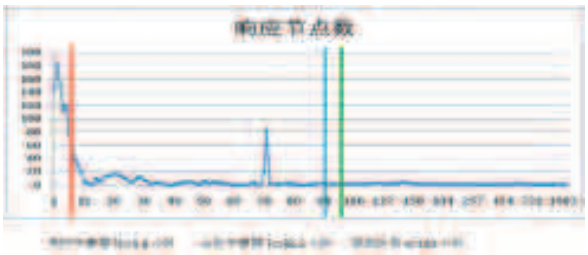

图 6：微博半衰期示意图

定义 8　半衰期趋势比：我们用历时半衰期的数值比上共时半衰期的数值，可以判断微博传播的阶段性趋势，我们将这个比值定义为微博信息传播的半衰期趋势比，表示为 $T/(T_n-T)$。

因为 $T_n>T$，因此这一比值为正数，数值越小，说明在观测时刻判断其传播趋势越弱，反之数值越大，说明其传播正在发展中。如果半衰期趋势比小于一定数量，说明观测时间距微博信息传播结束时刻较长。通常我们对半衰期趋势比大于 1/4 的微博，进行观测时刻的现时状态的研究，对于半衰期趋势比小于 1/4 的微博只进行历史传播过程的研究 [1]。

第二，微博半衰期观测的对象不是直接测量源微博本身的信息老化程度，而是通过对其再传播数量的时间分布的观测，间接对源微博的信息老化程度进行测量。

第三，在计算中响应节点数 c 和响应总节点数 C 与响应范围 U 不同 [2]，U 是记录不重复节点数，响应节点数 c 与 C 是不对再传播节点数去重的。

4.1.3.2 微博信息传播的结束

(1) 微博信息传播结束的概念

定义 9　微博信息传播结束：我们将微博信息传播过程中，响应为零的时间连续累积，与源微博起始至最后一个响应时刻的累积时长求得比值，这个比值大于等于微博信息传播的半衰期趋势比时，说明其不仅传播响应持续为零，而且其静默时间的长度与其传播发展的趋势是呼应的。这种情况我们将其定义为微博信息传播结束。为方便计算，我们通常将半衰期趋势比设为 1/4 。即响应为零的时间连续累积，大于等于源微博起始至最后一个响应时刻的累积时长的 1/4 时，视该微博传播过程阶段性结束，最后一个响应时刻，为本次观测时刻中可测量的传播结束时刻。

具体来讲，半衰期趋势比越小，说明传播力越弱，静默时长大于等于这一数值即可视为传播过程结束。反之，半衰期趋势比很大，说明其传播在快速发展中，其数值会越大，则静默状态要持续很长时间才可能在时长上大于其数值，这时才能判定其传播结束。这种定义方法，综合判断了静默时长和传播发展趋势，可以比较客观地反映微博信息传播结束这种情况。而根据对半衰期趋势比的统计 [3]，绝大部分微博半衰期趋势比都大于 4，因此在定义中使用 1/4 这个参数基本可以概括大多情况。如果需要严格判断个别微博的结束时刻，可以取半衰期趋势比的倒数进行计量。

(2) 微博信息传播周期时长

定义 10 微博信息传播周期：自源微博发布至传播结束的总时长为微博信息传播这个周期的时长，通常用小时和分钟数来表示。

(3) 微博信息传播结束和传播周期时长的数学定义

设源微博 X 自 t_0 时发布，在 t_1 秒内有 c_1 个响应进行了再传播，至观测时刻 t_n 秒内有 c_n 个响应进行了再传播，则当某一时刻 T′ 秒时 $c_{T'}=0$，有 $C_{T',n}=\sum_{i=T'}^{n}c_i=0$，可求出 T 为微博 X 的历时半衰期，$T_n-T$ 为微博 X 的共时半衰期，且有 $\frac{T_n-T'}{T'}\geq\frac{T}{T-T_n}$ 时，则在观测时刻可定义源微博 X 传播本周期结束，T′ 时刻为源微博 X 的传播可测量的结束时刻，而 $T'-t_0$ 为微博信息传播的周期时长。为方便实际使用，我们通常会以 $\frac{T_n-T'}{T'}\geq\frac{1}{4}$ 来进行简化计算。

4.1.4 微博信息传播半衰期的数据统计与分析

4.1.4.1　微博信息传播半衰期的计算方法

使用新浪微博提供的 API，随机抽取微博中的各项数据作为源微博数据样本库。再从源微博数据样本库中随机抽取 1 万、5 万和 10 万三组样本数据，选取其中有效数据，计算每一条微博信息的历时半衰期、共时半衰期以及半衰期趋势比。对半衰期各组数据的半衰期按升序排列，进

1 这里将区分现时状态研究和历史过程研究的微博半衰期趋势比的数值设定为 1/4，是与微博信息传播线束的定义相对应，即只对传播尚未结束的微博进行现时状态和趋势的研究是有意义的。
2 响应范围 U 的定义参见本文上节中第 5 项内容。
3 参见本文中 4.1.4　部分“微博信息传播半衰期的数据统计与分析”中的数据，在 1 万、5 万和 10 万的数据样本量的统计结果中，样本中 95% 的数据的半衰期趋势比的分位数均约 1/17000 以上，远远小于 1/4。

行通过 K-S 正态检验（Kolmogorov-Smirnov 检验）并对比分析历时半衰期的分布情况。确定统计数据范围，排除少量异常值。最后进行统计，求得微博信息传播的平均历时半衰期。分析半衰期趋势比，并验证微博信息传播结束的数学定义中参数 1/4 的合理性。

其中每批样本数据的历时半衰期的计算方法如下：

⑴从新浪提供的 API 接口中随机抽取需要的源微薄 ID 和源微博的创建时间，其创建时间记为 T_{i0}，i=1,…,N。

⑵提取每一个有效的源微博的所有转发微博的微博 ID 和转发时间，评价微博的微博 ID 和评价时间。将他们合并一起，记录时间为 T_{ij},j=1,…,N。N 为所有转发和评价以及评价的转发记录的总个数。

⑶计算并记录每一个源微博的时间差 T_{ij} - T_{i0}（单位秒）的值。

⑷将无转发且无评论的源微博数据，和新浪 API 提取中的出现错误的数据作为无效数据剔除。即剔除掉小于 0 的值，计算中值。该值即为这个微博的历时半衰期。

⑸将所有样本数据的历时半衰期数值进行升序排列，以每 2 个小时作为一个考察单位，并进行频次分布的分析。

⑹由于数据非正态分布的特点，需排除少量特殊数值的样本，确定有效统计范围。

⑺在有效统计的范围内，以秒为单位计算所有微博的平均历时半衰期和微博半衰期趋势比的95%的分位数。

4.1.4.2 历时半衰期和半衰期趋势比的统计

（1）历时半衰期对数正态分布检验

根据上述微博历时半衰期的计算方法，计算出所有有效数据样本的源微博历时半衰期。在每组数据中，将每 2 个小时作为一个单位，对历时半衰期数值进行升序排列。绘制历时半衰期对数的直方图，进行数据分布的观察。从图中可以看出，历时半衰期的对数有近似服从正态分布的趋势，并进行正态分布检验。以下以 5 万数据样本组的检验数据为例：

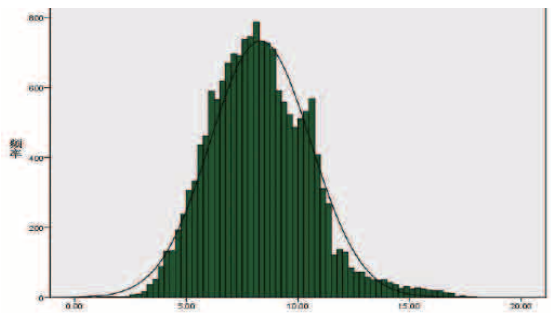

图 7 : 历时半衰期对数的直方图

通过 Kolmogorov-Smirnov 正态检验，根据检验结果 p<0.05，可知微博历时半衰期的分布不服从对数正态分布。

表 4 : Kolmogorov-Smirnov 正态检验

	Kolmogorov-Smirnova		
	统计量	df	Sig.
半衰期对数	.020	16447	.000

a. Lilliefors 显著水平修正

（2）排除异常数据确定统计范围

根据正态分布检验的结果，平均历时半衰期的计算需排除异常数值，以免影响统计结果。从微博历时半衰期异常值对平均值统计的影响图中可见，统计数据范围的扩大，数值较大的异常值会使平均历时半衰期的统计数值偏大，甚至严重偏离实际情况。

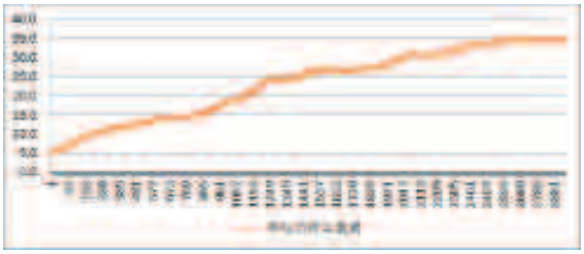

图 8 : 微博历时半衰期异常值对平均值统计的影响图

为确定合理的统计范围，我们需通过异常数据分组比较，来找到异常值影响增大的拐点。对 1 万、5 万和 10 万三个数据样本组的数据，分别进行异常值分组求平均值进行比较。将统计数据范围从历时半衰期值小于等于 24 小时的数据开始，每增长 2 个小时的数据量进行一次平均值的统计，直到历时半衰期数值小于等于 72 小时的数据全部统计在内。我们明显看到这是一个连续上升的数据序列，并无异常值造成统计值陡然增大的拐点。

表 5 : 分组数据不同统计范围平均历时半衰期

统计数据范围	平时历时半衰期统计值 (小时)		
	1 万数据样本组	5 万数据样本组	10 万数据样本组
24 小时	2.758	2.899	2.904
26 小时	2.882	3.016	3.027
28 小时	2.955	3.082	3.112
30 小时	2.999	3.121	3.189
32 小时	3.063	3.172	3.240
34 小时	3.121	3.257	3.298
36 小时	3.271	3.304	3.376
38 小时	3.299	3.369	3.445
40 小时	3.426	3.462	3.502
42 小时	3.488	3.512	3.559

44 小时	3.532	3.570	3.612
46 小时	3.565	3.623	3.677
48 小时	3.661	3.721	3.770
50 小时	3.735	3.810	3.844
52 小时	3.774	3.879	3.899
54 小时	3.787	3.922	3.948
56 小时	3.844	3.958	3.983
58 小时	3.916	3.985	4.032
60 小时	3.991	4.012	4.091
62 小时	4.069	4.034	4.138
64 小时	4.085	4.052	4.179
66 小时	4.085	4.083	4.204
68 小时	4.119	4.154	4.266
70 小时	4.137	4.211	4.317
72 小时	4.210	4.274	4.390

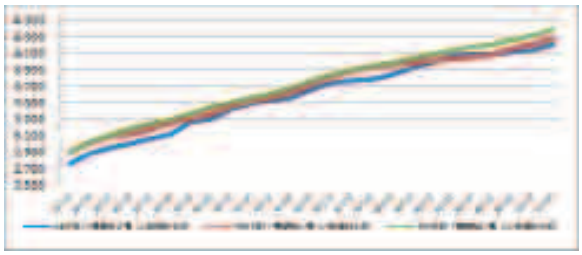

图 9 : 各样本组不同数据统计范围对平均历时半衰期统计值的影响

对微博历时半衰期统计数据对比表的分析可见，到 46-48 小时这组数据时，三组统计数据的累积百分比都在 95% 左右。并且三组统计数据中,从 24-26 小时数据组之后，每组数据的有效百分比都小于 0.5%。即在去除了历时半衰期在 48 小时以上的数据后，统计总量上仍然覆盖了绝大部分数据，去除的后续各组数据的增量都非常小，但对最终统计结果的影响却比较大。

表 6 : 不同数据样本的历时半衰期分布对比表

	1万样本数据统计			5万样本数据统计			10万样本数据统计		
	频率	有效 %	累积 %	频率	有效 %	累积 %	频率	有效%	累积 %
2 小时以内	2423	63.1	63.1	10404	62.7	62.7	20594	62.3	62.3
2-4 小时	381	9.9	73.1	1601	9.6	72.3	3284	9.9	72.3
4-6 小时	192	5	78.1	834	5	77.4	1587	4.8	77.1
6-8 小时	145	3.8	81.9	584	3.5	80.9	1113	3.4	80.4
8-10 小时	106	2.8	84.6	476	2.9	83.8	1015	3.1	83.5
10-12 小时	84	2.2	86.8	436	2.6	86.4	815	2.5	86
12-14 小时	63	1.6	88.5	303	1.8	88.2	613	1.9	87.8
14-16 小时	32	0.8	89.3	193	1.2	89.4	402	1.2	89.1
16-18 小时	26	0.7	90	173	1	90.4	318	1	90
18-20 小时	27	0.7	90.7	123	0.7	91.2	260	0.8	90.8
20-22 小时	28	0.7	91.4	121	0.7	91.9	255	0.8	91.6
22-24 小时	33	0.9	92.3	109	0.7	92.6	221	0.7	92.3
24-26 小时	20	0.5	92.8	82	0.5	93	171	0.5	92.8
26-28 小时	11	0.3	93.1	43	0.3	93.3	109	0.3	93.1
28-30 小时	6	0.2	93.2	24	0.1	93.4	93	0.3	93.4
30-32 小时	8	0.2	93.4	28	0.2	93.6	56	0.2	93.5
32-34 小时	7	0.2	93.6	45	0.3	93.9	61	0.2	93.7
34-36 小时	17	0.4	94.1	23	0.1	94	76	0.2	94
36-38 小时	3	0.1	94.1	30	0.2	94.2	64	0.2	94.2
38-40 小时	13	0.3	94.5	41	0.2	94.5	50	0.2	94.3
40-42 小时	6	0.2	94.6	21	0.1	94.6	47	0.1	94.5
42-44 小时	4	0.1	94.7	23	0.1	94.7	42	0.1	94.6
44-46 小时	3	0.1	94.8	20	0.1	94.8	49	0.1	94.7
46-48 小时	8	0.2	95	36	0.2	95.1	68	0.2	94.9
48 小时以上	191	5	100	820	4.9	100	1674	5.1	100
合计	3837	100	100	16593	100	100	33037	100	100

因此我们去除这部分的异常值，以微博历时半衰期小于等于 48 小时的数据来求出平均历时半衰期的数值。

(3) 1 万数据样本的统计结果

1 万数据样本统计出的平均历时半衰期约 3.66 小时，半衰期趋势比的 95% 分位数为 1/177465。（去除半衰期在 48 小时以上数据样本，使用数据占总数据的 95.0%）

表 7 : 1 万数据样本的历时半衰期分布

数据分组	频率	百分比	有效百分比	累积百分比
2 小时以内	2423	63.1	63.1	63.1
2-4 小时	381	9.9	9.9	73.1
4-6 小时	192	5.0	5.0	78.1
6-8 小时	145	3.8	3.8	81.9
8-10 小时	106	2.8	2.8	84.6
10-12 小时	84	2.2	2.2	86.8
12-14 小时	63	1.6	1.6	88.5
14-16 小时	32	.8	.8	89.3
16-18 小时	26	.7	.7	90.0
18-20 小时	27	.7	.7	90.7
20-22 小时	28	.7	.7	91.4
22-24 小时	33	.9	.9	92.3
24-26 小时	20	.5	.5	92.8
26-28 小时	11	.3	.3	93.1
28-30 小时	6	.2	.2	93.2
30-32 小时	8	.2	.2	93.4

32-34 小时	7	.2	.2	93.6
34-36 小时	17	.4	.4	94.1
36-38 小时	3	.1	.1	94.1
38-40 小时	13	.3	.3	94.5
40-42 小时	6	.2	.2	94.6
42-44 小时	4	.1	.1	94.7
44-46 小时	3	.1	.1	94.8
46-48 小时	8	.2	.2	95.0
48 小时以上	191	5.0	5.0	100.0
合计	3837	100.0	100.0	

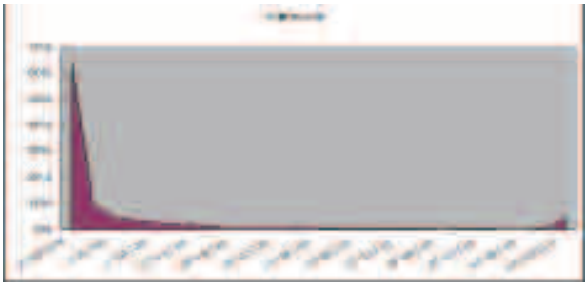

图 10 : 1 万数据样本的历时半衰期频次分布图

（4）5 万数据样本的统计结果

5 万数据样本统计出的平均历时半衰期约 3.72 小时，半衰期趋势比的 95% 分位数为 1/182231。（去除半衰期在 48 小时以上数据样本，使用数据占总数据的 95.1%）

表 8： 5 万数据样本的历时半衰期分布

数据分组	频率	百分比	有效百分比	累积百分比
2 小时以内	10404	62.7	62.7	62.7
2-4 小时	1601	9.6	9.6	72.3
4-6 小时	834	5.0	5.0	77.4
6-8 小时	584	3.5	3.5	80.9
8-10 小时	476	2.9	2.9	83.8
10-12 小时	436	2.6	2.6	86.4
12-14 小时	303	1.8	1.8	88.2
14-16 小时	193	1.2	1.2	89.4
16-18 小时	173	1.0	1.0	90.4
18-20 小时	123	.7	.7	91.2
20-22 小时	121	.7	.7	91.9
22-24 小时	109	.7	.7	92.6
24-26 小时	82	.5	.5	93.0
26-28 小时	43	.3	.3	93.3
28-30 小时	24	.1	.1	93.4
30-32 小时	28	.2	.2	93.6
32-34 小时	45	.3	.3	93.9
34-36 小时	23	.1	.1	94.0
36-38 小时	30	.2	.2	94.2
38-40 小时	41	.2	.2	94.5
40-42 小时	21	.1	.1	94.6
42-44 小时	23	.1	.1	94.7
44-46 小时	20	.1	.1	94.8
46-48 小时	36	.2	.2	95.1
48 小时以上	820	4.9	4.9	100.0
合计	16593	100.0	100.0	

图 11 : 5 万数据样本的历时半衰期频次分布图

（5）10 万数据样本的统计结果

5 万数据样本统计出的平均历时半衰期约 3.77 小时，半衰期趋势比的 95% 分位数为 1/189426。（去除半衰期在 48 小时以上数据样本，使用数据占总数据的 94.9%）

表 9 : 10 万数据样本的历时半衰期分布

数据分组	频率	百分比	有效百分比	累积百分比
2 小时以内	20594	62.3	62.3	62.3
2-4 小时	3284	9.9	9.9	72.3
4-6 小时	1587	4.8	4.8	77.1
6-8 小时	1113	3.4	3.4	80.4
8-10 小时	1015	3.1	3.1	83.5
10-12 小时	815	2.5	2.5	86.0
12-14 小时	613	1.9	1.9	87.8
14-16 小时	402	1.2	1.2	89.1
16-18 小时	318	1.0	1.0	90.0
18-20 小时	260	.8	.8	90.8
20-22 小时	255	.8	.8	91.6
22-24 小时	221	.7	.7	92.3
24-26 小时	171	.5	.5	92.8
26-28 小时	109	.3	.3	93.1
28-30 小时	93	.3	.3	93.4
30-32 小时	56	.2	.2	93.5
32-34 小时	61	.2	.2	93.7

34-36 小时	76	.2	.2	94.0
36-38 小时	64	.2	.2	94.2
38-40 小时	50	.2	.2	94.3
40-42 小时	47	.1	.1	94.
42-44 小时	42	.1	.1	94.6
44-46 小时	49	.1	.1	94.7
46-48 小时	68	.2	.2	94.9
48 小时以上	1674	5.1	5.1	100.0
合计	33037	100.0	100.0	

图 12 : 10 万数据样本的历时半衰期频次分布图

4.1.4.3 历时半衰期统计结果讨论

计算结果可知，目前新浪微博系统中的微博信息传播的平均历时半衰期约为 3.7 小时。

表 10 : 三组统计结果对比表

	平均历时半衰期（小时）	半衰期趋势比
1 万数据	3.66	1/177465
5 万数据	3.72	1/182231
10 万数据	3.77	1/189426

根据上述三个样本库的数据统计，半衰期趋势比分别为 1/177465 和 1/182231 以及 1/189426 都远远小于 1/4，因此微博信息传播结束定义中，使用 1/4 这个简化计算的参数是远大于统计的平均值，基本不会造成误判。（数值非常小的原因是，统计样本中历史数据较多，造成共时半衰期非常大，进而影响半衰期趋势比。）如需准确判断可使用本文中给出的准确数据公式。

根据对微博信息传播的平均半衰期和半衰期趋势比的测定以及微博信息传播结束的数学定义，我们可以在任一观测时刻，对任何一条源微博进行历时半衰期、共时半衰期的测量，计算半衰期趋势比。对未结束传播的源微博，所处传的播阶段进行判断。对已结束传播的源微博的传播周期的长度进行计算。因此可以对微博传播持续时间进行排名分析，并综合其它数据观测微博信息传播的一般性规律。同时，如观测时刻有一条源微博的历时半衰期比较大，即大于 3.7 这个平均微博历时半衰期，且半衰期趋势比非常大，则可判断这条源微博的传播周期将会比较长，影响相对比较大。我们进而可以综合其它因素，对正在传播中的微博的影响力进行有效预测。

4.2 微博影响力节点的数学模型

4.2.1 微博的传播力简述

微博的传播力量到底有多大？

新浪 CEO 曹国伟透露，根据他们的已有数据，一条微博经过三个节点的传播，最大可能在整个微博用户群中实现 97.3%的覆盖率[1]。20 世纪 60 年代，由美国心理学家米尔格伦提出了著名的“六度空间”理论，又称作六度分隔 Six Degrees of Separation 理论。这个理论可以通俗地阐述为：“你和任何一个陌生人之间所间隔的人不会超过六个，也就是说，最多通过六个人你就能够认识任何一个陌生人。”如果把六个人看作六个传播节点，通过微博，只需三个节点，就可以覆盖 97.3%的六度空间。微博传播力的广度可见一斑。

杜子健在《微力无边》一书中认为[24]，微博的工作原理几乎和核反应堆一模一样：“当铀 235 的原子核受到外来中子轰击时，一个原子核会吸收一个中子分裂成两个质量较小的原子核，同时放出 2 ～ 3 个中子。裂变产生的中子又去轰击另外的铀 235 原子核……”。他以“宜黄事件”为例，中子“@ 钟如九”在微博上遭遇（轰击）铀“@ 北京厨子”后迅速分裂出两个或多个小原子核，同时又释放（转发）2 ～ 3 个中子，裂变出的中子再去轰击铀“@ 王小山”引起新裂变……，于是携带巨大核能的“链式发应”开始发威。杜子健说：“微博，是社会信息反应堆，是个体信息和大众情绪发生碰撞以后可能产生巨大核变能量的情绪反映堆。”

4.2.2 微博传播的内在机制

核反应堆的比喻形象表现了微博传播的引爆力度。喻国明所著《微博：一种新传播形态的考察》一书中，精细地剖析了微博传播的内在机制，他将一条微博传播轨迹中，担当不同角色的节点分为[25]：

1 数据来自：2011 年 4 月 27 日，新浪 CEO 曹国伟在北京国家会议中心举行的第三届全球移动互联网大会（GMIC2011）上的演讲。

1) 核心节点：信息的源头，其他用户关注的核心。

2) 桥节点：核心用户传播信息的扩散者。

3) 长尾节点：借助桥节点的作用才能接触到核心节点的信息。他们是沉默的大多数，单个发言的影响力小，但集合起来却发挥巨大的能量。

无疑，桥节点在微博的整个传播过程中，起着举足轻重般的枢纽作用。那么有没有可能，用数学公式，对于桥节点进行量化计算，使得我们对于桥节点的传播贡献有一个具体量度，进而推算出那些最具影响力的关键节点？

4.2.3 微博影响力节点运算的数学模型

微博传播节点，是整个传播链中的一个环节，所以它提供的传播通道、轨迹的形式、质量直接影响了它能所产生的传播效应。同时，它的传播通道，在微博体系的拓扑结构中，是以节点之间的关注、转发、评论来表征的。关注、转发、评论，本身是可以量化为数字的，所以微博某节点的被关注数量，所传播消息的转发、评论数，好比民主投票，数值越高其节点的传播价值越高。

这种通过信息单元节点的互投票方式，来决定信息单元的价值，让我们想起了 Google 著名的 Page Rank 算法[26]：

$$R(u) = c\sum_{v\subseteq B(u)}\frac{R(v)}{N(v)} \quad (1)$$

其中：

- R(v)：页面 v 的 Page Rank 值。
- N(v)：从页面 v 链接到其它页面的链接数目。
- B(u)：链入页面 u 的链接数目。
- C：规范化因子。

所以，如果将网页的互链行为，转化为微博节点之间的转发、评论，我们可以略对 Page Rank 算法进行修正，变为微博影响力的计算公式：

$$TipPoint(u) = c\sum_{v\subseteq CB(u)} TipPoint(v)\,\mathrm{F} \quad (2)$$

其中：

- TipPoint(v)：微博节点 V 的传播影响力。
- c：阻尼系数。
- CB(u)：用户 U 的微博被转发、评论的数量（转发和评论权值假定相等）。
- F：权值分配因子，初使值设置为 1。

以一个现实中发生的微博为例，（抓取时间：2012 年 2 月 22 日）如下文所示：

方舟子 ：路透社报道：针对《我在美国当市长助理》一书中绘声绘色描述如何受纽黑文市市长 DeStefano 指派调查处理两名警察超假三天一事，市政府发言人 Benton 说，DeStefano 市长对此事毫无印象，市长也绝不会派一名实习生去处理本该由警察局内务部去处理的事务。

转发 (578)| 收藏 | 评论 (448) 今天 08:30 来自新浪微博

用上面的引爆点公式，计算出上述微博传播过程中，其影响力最大的 50 个传播节点如表 11 所示。

表 11：F=1 时前 50 个影响力节点以及作者信息

序号	影响力	作者姓名	粉丝数	是否认证	所发微博数	与博主关系
1	73	LonelyPlanet	78299	是	2434	0
2	38	麦韬旅游	18072	是	11144	3
3	34	南方日报	128994	是	10294	0
4	29	petriv	3286	否	8469	0
5	26	浙江省旅游局	822241	是	4096	0
6	24	凯撒旅游的官方微博	589155	是	2670	0
7	19	宝中堂	10314	是	3239	0
8	15	曹增辉	32682	否	17891	0
9	15	新浪尚品	298291	是	5469	2
10	9	北京新闻广播	393545	是	15817	0
11	8	元小挞	41	否	111	0
12	6	徐香菜	425	否	2123	0
13	6	中青旅百变自由行	79250	是	2678	1
14	5	李清	5481	是	5866	0
15	5	外交小灵通_智囊团	15050	是	2364	0
16	4	可乐泡面	1778	否	3762	0
17	4	汴人郭威	2480	否	10217	0
18	4	甘肃刘维忠	507595	是	5574	0
19	3	环海花园	2557	否	3832	0
20	3	中青报周凯	6815	是	19047	0
21	3	陈强微博	15597	否	13319	0
22	3	琪缘	33300	是	2975	0
23	3	孟湛	29026	是	681	0
24	3	亚洲廉航网	8855	是	1164	1

25	2	马尔代夫 Cosmore	2688	否	4082	3
26	2	MaldivesLily	490	否	424	0
27	2	爱小澈	681	是	2537	1
28	2	中国康辉旅行社一北京	1962	是	170	1
29	2	岐岐在斯里兰卡的日子	757	否	1876	0
30	2	严海锋	1534	是	1947	0
31	2	巫云峰	12950	是	7115	0
32	2	广州碎片	488	否	574	1
33	2	上海市旅游局张卉	1530	是	2851	0
34	2	辉摄轨迹	505	否	677	1
35	2	游多多	19257	是	3796	1
36	2	枉硼緗存	1350	否	5047	0
37	1	椿椏的家园	72	否	595	0
38	1	Fleur_Sauvage	93	否	906	0
39	1	83 年的胖小猪楠楠	184	否	571	0
40	1	Viiiiiii 妞	115	否	471	0
41	1	佳佳奶糖 07	108	否	1792	0
42	1	阿昕 -- 爱茵假期	1	否	1	0
43	1	郑州升龙广场	110	是	100	0
44	1	姚楼楼姚	14	否	77	0
45	1	伏地豚	70	否	285	0
46	1	论雯 yoki_ 马代代理	4061	否	435	1
47	1	旦小乙	197	否	1419	0
48	1	张岩空间	304	是	267	0
49	1	米童世界	38	否	591	0
50	1	超级爬爬凳	1	否	26	0

注：在“与搏主关系”列表中，0——两者没有任何关系；1——评论者关注了源微博作者，源微博作者没有关注评论者；2——评论者没有源微博作者，源微博作者关注了评论者；3——两者是互粉关系。

如果 F 取 1，无疑默认每一个对源微博转发、评论的节点，其信息效用是相等的，凭常识就可以判断，相同的信息，被不同的传播主题表达，其效用是不同的。比如一个社会口号，由大众明星表达，还是让某个街头小孩喊几声，其社会效用天壤之别。所在这里，我们先根据每个节点的粉丝数（粉丝越多，传播覆盖面越大）来修正它的传播权重。所以令：

$$F=\frac{A_j}{\sum_{j=1}^{CB(u)}A_j} \quad (3)$$

A_j：节点 V 的粉丝数。

即：

$$TipPoint(u)=c\sum_{i=1}^{CB(u)}\frac{A_j}{\sum_{j=1}^{CB(u)}A_j}TipPoint(v) \quad (4)$$

4.2.4 未来的研究规划

- 引爆点数学公式，将考虑引进用户活跃度、用户认证等因素进行修正。
- 扩大用于引爆点计算的微薄样本数量，不仅探索影响力节点和它自身属性（比如粉丝数、用户活跃度、是否名人或认证、与源微博博主的关系）的各种潜在联系；而且进一步分析引爆点集群的各种特征（比如拓扑结构等）。
- 与醒客工场关于微博传播半衰期研究进行结合，分析微博传播过程中，发生传播拐点时，其传播节点的特征，进一步修正引爆点数学公式。

4.3 微博传播的心理学研究模型

4.3.1 研究目标

构建微博的情绪语料库和情绪层次模型，编制微博使用偏好问卷，探讨影响微博传播的心理学因素。进而发现微博的传播规律，将其广泛运用于产品推广、广告营销，股市预测和电影票房预测，微博谣言的早期预警、网络舆情监测与引导，突发公共卫生事件的应对以及建构国民幸福感或城市幸福感指标等领域。

4.3.2 研究思路

Prasad（1950）[27] 强调多水平方法（multi-level approach）在研究中的重要性，即需要同时探讨个体的心理因素和团体的影响。因此，基于心理学视角，本研究采用多水平方法，从个体心理和群体影响两个方面来探讨微博的传播规律。个体心理的研究主要回答“什么样的个体更容易传播微博”的问题，主要包括个体使用微博的动机以及个体的整体人格和特定人格特质。群体影响的研究主要回答“群体性心理学指标怎么来预测微博传播过程”的问题，主要涉及到公众情绪。

4.3.3 研究的内容

本研究主要有以下几个研究内容：（1）微博情绪语料库的构建；（2）微博使用偏好问卷的编制；（3）微博传播过程中群体影响的研究；（4）微博传播过程中个体差

异的研究。

4.3.4 研究方法

本研究主要的研究方法是基于微博数据的分析，具体通过微博数据的挖掘，构建相应的数学模型或绘制曲线图；还有问卷调查法，主要编制微博使用偏好问卷。具体针对不同的研究内容采用不同的研究方法。

（1）微博情绪语料库的构建。基于微博抓取的数据，从微博琐碎的语言表达中提炼情绪词汇。根据这些情绪词汇构建情绪的层次模型，根据粒度不同模型分成三层，最底层将情绪细分为 9 类，包括高兴、喜爱、惊奇、愤怒、悲伤、害怕、厌恶、害羞和中立；中间一层将情绪分成积极情绪、消极情绪和中立情绪；最顶层将情绪分成倾向性情绪和中立性情绪（见图 13）。

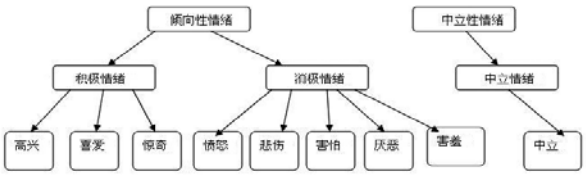

图 13：情绪的层次模型

（2）公众情绪是否和微博传播过程有关。根据微博内容社交属性的分类体系，即将微博内容分成信息咨询和评论、心理（情感）诉求，号召行动（个人、群体）三大类，分别选择这三大类的微博，探讨在这三个类别的微博传播过程中公众情绪是如何变化。根据情绪的层次模型对微博评论进行情绪分类，绘制第一层、第二层和第三层情绪分类的时间变化坡度图。具体包括“整体情绪指标”和“单一情绪指标”。

整体情绪指标包括：

a. 情绪分布度，是刻画不同情绪的分布情况，参照刘志明和刘鲁（2010）[28] 的公式，该公式定义如下：

$$D_t = \sum_i (N_{t,i} - E_t)^2 it \tag{5}$$

其中，$N_{t,i}$ 表示 t 时刻第 i 种情绪对应的评论数量，E_t 表示 t 时刻不同类别情绪对应的评论数量的均值，使用方差来表示情绪的分布情况。

b. 情绪分布变化率，用来刻画情绪分布的变化情况，参照刘志明和刘鲁 [28] 的公式，该公式定义如下：

$$R = \left| \frac{D_j - D_i}{T_j - T_i} \right| \tag{6}$$

其中，D_i，D_j 分别表示 i，j 时刻情绪的分布度，T_i，T_j 分别表示时刻 i，j。

单一情绪指标包括：

a. 情绪热度，用来刻画某种情绪在所有评论中所占的比例，参照刘志明和刘鲁（2010）[28] 的公式，该公式定义如下：

$$H_{t,i} = \frac{N_{t,i}}{\sum_i N_{t,i}} \tag{7}$$

其中，$N_{t,i}$ 表示 t 时刻第 i 种情绪对应的评论数量。

b. 情绪拐点，用来刻画某种情绪的变化情况，参照刘志明和刘鲁（2010）[28] 的公式，该公式定义如下：

$$G_k = \frac{N_{k,j} - N_{k,i}}{T_j - T_i} \tag{8}$$

其中，$N_{k,i}$，$N_{k,j}$ 表示时刻 i，j 第 k 种情绪对应的评论数量，T_i，T_j 分别表示时刻 i，j。

（3）编制微博使用偏好问卷。用户在使用微博时由于动机和需要不同，他们经常使用的微博服务或功能的内容也有所不同。如可能存在社交服务、信息服务和娱乐服务等。首先通过文献调研、访谈法和小范围人群测试，获得微博使用偏好问卷的初始题目，然后通过选取有代表性的大样本，对初始问卷进行修订，经过对其心理测量学指标的检验，最终形成信度和效度较好的正式微博使用偏好问卷。

（4）微博传播过程中的个体差异。采用问卷调查法，测试内容包括三大块。第一大块是被试的基本人口学资料，如性别、年龄、受教育程度、职业、收入等。第二大块是心理学量表，包括微博使用偏好问卷、大五人格量表（包括神经质、内外向、开放性、尽责性和宜人性，用来测试被试的人格全貌）和无法忍受不确定性问卷（某种特定的人格特质，用于测量被试对不确定性情景的忍受程度）。第三大块是微博传播特征指标，如被试的微博使用频率等。通过多元回归分析，用基本人口学变量和心理学变量来共同预测微博使用频率，以发现影响微博传播过程的重要个体变量。

5 微博信息的应用统计分析

5.1 新浪微博数据的相关基础分析

样本数据：随机抽取的 50000 条微博。本部分内容分

为两个部分。第一部分基本分析旨在详细地描述样本数据在时间发布时间、内容形式、客户端、响应量和响应时间方面的分布特征，从而对样本的基本规律有一个了解和描述。第二部分回归分析旨在探讨发布时间、内容形式、客户端、微博作者等因素对微博响应量（包括评论和转发）的影响。

5.1.1 基本分析

5.1.1.1 微博发布时间

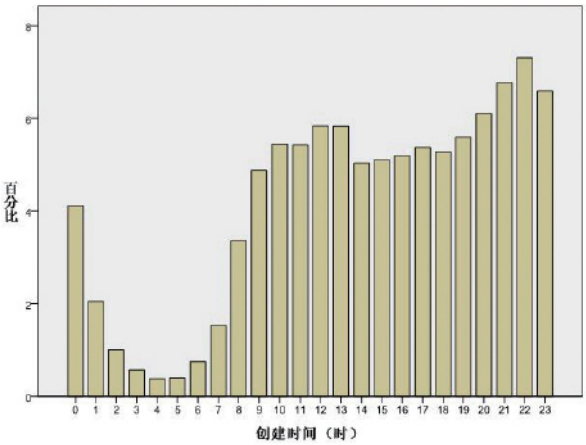

图 14 : 源微博创建时间

图 14 中，x 轴为源微博的创建时间，按 24 小时计算。y 轴为每个时段发布微博所占微博总数的百分比。

由上图可见，从凌晨 1 点到 7 点是微博发布频次较低的时段。从早晨 8 点开始呈迅速上升趋势，到中午 12、13 点左右达到一个小高峰，下午则一直保持稳定的较高频次，晚上 19 时后频次增加，到 22 点达到最高峰值，之后两个小时迅速下降。这基本上反映了微博用户的作息规律，下一步的分析将区分个人用户和组织用户，区分工作日和周末。

5.1.1.2 微博内容形式

表 12 : 微博内容形式

	@	表情符	URL	#
频次	19773	11963	3687	1807
百分比	33%	19.9%	6.1%	3%

对微博内容的形式进行分析显示，大约 1/3 的微博是有方向性的，即包含 @ 符。约有 1/5 的微博使用表情符号配合文字进行表达。引用 URL 的占 6% 左右，而使用 Hashtag ＃来标记主题的则占 3%。

5.1.1.3 微博发布客户端

大约 60%的微博是通过 weibo.com 网站发布的。而其它诸如手机和平板电脑等移动设备客户端累计也占了一定的比例。

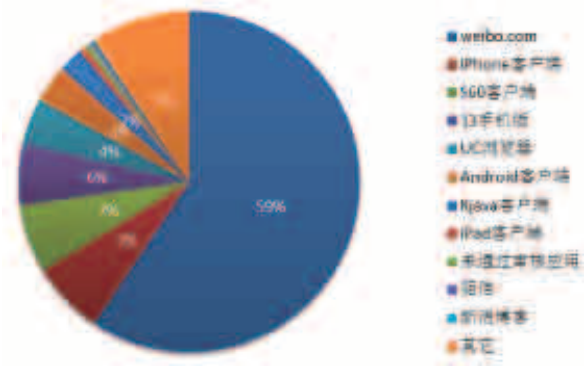

图 15 : 发布微博使用的客户端

5.1.1.4 微博响应量

表 13 : 微博转发频次

转发次数	频次	百分比
无转发	49401	82.3%
1 次	7259	12.1%
2 次	1771	3.0%
3 次	603	1.0%
4 次及以上	966	1.6%
合计	60000	100.0%

表 14 : 微博评论频次

评论次数	频次	百分比
无评论	46134	76.9%
1 次	4764	7.9%
2 次	2854	4.8%
3 次	1420	2.4%
4 次	1110	1.9%
5 次	719	1.2%
6 次	582	1.0%
7 次	415	0.7%
8 次	366	0.6%
9 次	266	0.4%
10 次	207	0.3%
11 次及以上	1163	1.9%
合计	60000	100.0%

表 13 和 14 显示，约占 80%左右的微博没有获得任何形式的响应（转发或评论）。对于没有响应的源微博来说，微博作者的粉丝有可能接收到此微博，但并没有在微

博上对该微博进行传播。这些微博的传播范围小于等于微博作者的粉丝数量。

5.1.1.5 响应时间

表 15 : 转发的时间分布

时段	第一次转发		最后一次转发	
	频次	累积百分比	频次	累计百分比
1 小时以内	3846	64.4%	2875	48.2%
1 小时到 1 天	1673	92.5%	2313	86.9%
1 天到 1 月	325	97.9%	578	96.6%
1 月到 6 月	108	99.7%	173	99.5%
6 月到 1 年	15	100.0%	28	100.0%
1 年到两年	2	100.0%	2	100.0%

表 15 分别显示了微博的第一次和最后一次转发的时间分布。在一小时之内，约有 64%的微博获得了第一次转发，而大约一半（48.2%）的微博完成了所有转发。在一天内大部分微博获得其首次转发（92.5%）和最后一次转发 (86.9%)。一个月内绝大多数微博（96.6%）完成了其所有的转发。

表 16 : 评论的时间分布

时段	第一次评论		最后一次评论	
	频次	累积百分比	频次	累计百分比
1 小时以内	9435	69.2	4985	36.5
1 小时到 1 天	3619	95.7	6808	86.4
1 天到 1 月	504	99.4	1639	98.5
1 月到 6 月	70	99.9	175	99.7
6 月到 1 年	12	100.0	32	100.0
1 年到两年	3	100.0	4	100.0

表 16 分别显示了微博的第一次和最后一次评论的时间分布。除了在一小时之内完成最后一次评论的比例（36.5%）略少于在一小时之内完成最后一次转发的比例（48.2%）之外，其它的时间分布与表 15 显示了极为相似的规律。这表明使用转发和评论来作为微博响应时间的量度是基本可以通用的。

将表 15 和表 16 合并，产生了微博总响应时间分布表，即表 17。在 1 小时内，大约 70%的微博得到了第一次响应，38%的微博所有响应结束。1 天内，95%的微博获得第一次响应，86%的微博完成所有响应。一个月内，绝大多数微博响应完结。

表 17 : 微博响应的时间分布

时段	第一次响应		最后一次响应	
	频次	累积百分比	频次	累计百分比
1 小时以内	13281	69.6%	6279	37.8%
1 小时到 1 天	4255	95.3%	7967	85.8%
1 天到 1 月	632	99.1%	1977	97.8%
1 月到 6 月	131	99.9%	313	99.6%
6 月到 1 年	19	100.0%	53	100.0%
1 年到两年	5	100.0%	6	100.0%

5.1.2 回归分析

为分析各种因素对微博引起的响应的影响，我们进行了线性回归和 logistic 分析。因变量为一个微博的转发数、评论数，和总响应数（评论＋转发），自变量分为内容、形式、环境、作者四个组，每个组的变量和说明如表 18 所示：

表 18 : 变量说明

类别	自变量名	说明
内容	长度	微博内容的字数
形式	纯文本	微博是纯文本形式，不带有其它任何形式［0，1］
	@	微博内容带 @ 标记［0，1］
	＃＃主题	微博内容带用＃＃标记的主题[0,1]
	URL	微博内容带 URL 链接［0，1］
	表情	微博内容带表情符号［0，1］
环境	时间	微博的发布时间
作者	是否认证	作者是否为认证用户
	注册时间	作者的注册时间
	粉丝数	作者的粉丝数
	微博数	作者已发的微博总数

分析表明总响应数与粉丝数、互粉数、已发微博数以及微博内容的长度的关系不是线性关系，除去一些特殊点外，总响应数分别在由粉丝数、互粉数、已发微博数确定的一个区间里面。以粉丝数为例，在随着粉丝数的增大，总响应数也会相应地增大，但是粉丝数达到一定程度，总响应数在一定条件下会下降。

因此，我们进行了 logistic 回归分析。分析一条微博有没有响应，1 代表有相应，0 代表无响应，方法是做二分类 logistic 回归，默认情况下，即将出现频次最高的作为预测结果，结果的准确性为 67.5%。

5.1.2.1 基本结果

表 19 : 分类表 a,b

已观测		已预测		
		总响应数逻辑		百分比校正
步骤 0	总响应数逻辑 .00	.00	1.00	
	1.00	40494	0	100.0
总计百分比		19488	0	.0
				67.5

a. 模型中包括常量。 b. 切割值为 .500

如果将粉丝数、关注数、互粉数、已发微博数、用户、链接 URL、主题、包含表情、微博内容长度、认证作为自变量，结果的准确性为 68.6%。说明新变量的引入使预测的准确度提高 1.1%。提高的效果不是很明显，但也说明这些变量可以引起微博相应的变化，有待进一步考虑其它的因素。

表 20 : 分类表 a

已观测		已预测		
		总响应数逻辑		百分比校正
		无响应	有响应	
步骤 1	总响应数逻辑 无响应	38748	1746	95.7
	有响应	17108	2380	12.2
总计百分比				68.6

a. 切割值为 .500

5.1.2.2 变量的统计学意义

无 @ 用户微博获得响应的能力是有 @ 用户的 0.805 倍，有 @ 用户的微博更容易获得响应；无链接 URL 微博获得响应的能力是有链接 URL 的的 0.345 倍，有链接 URL 的更容易有响应；无主题微博获得响应的能力是有主题的 0.833 倍，有主题的更容易有响应；作者非认证的微博获得响应的能力是认证的 0.576 倍，有认证的微博更容易有响应；不包含表情的微博获得响应的能力是包含表情的 1.333 倍；不含表情的更容易响应；粉丝数、关注数、互粉数、已发微博数、微博内容长度分别增加一个单位值，比数自然对数值增加 1，0.999，1.003，1，1.007。

所有变量的 Wald 检验结果都小于 0.001，说明所有的选入变量都对微博的响应产生影响，具有统计学意义。

表 21 : 方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)
步骤 1a	粉丝数	.000	.000	26.405	1	.000	1.000
	关注数	-.001	.000	299.808	1	.000	.999
	互粉数	.003	.000	1089.798	1	.000	1.003
	已发微博数	.000	.000	22.488	1	.000	1.000
	用户	-.216	.021	101.211	1	.000	.805
	链接 URL	-1.065	.047	506.325	1	.000	.345
	主题	-.182	.058	9.961	1	.002	.833
	包含表情	.287	.022	163.729	1	.000	1.333
	微博内容长度	.007	.000	676.975	1	.000	1.007
	认证 (1)	-.552	.054	106.083	1	.000	.576
	常量	-.505	.055	83.418	1	.000	.603

a. 在步骤 1 中输入的变量：粉丝数，关注数，互粉数，已发微博数，用户，链接 URL，主题，，包含表情，微博内容长度，认证。

5.1.2.3 模型的拟合优度检验

使用 Hosmer- Lemeshow 检验，Hosmer- Lemeshow 统计量为 1096.254，自由度为 8，P 值 <0.05，模型拟合效果不佳。这说明当前自变量不足以完全解释因变量“微博响应是否”。有待进一步考察其它变量，特别是微博内容的影响。

表 22 : Hosmer 和 Lemeshow 检验

步骤	卡方	Df	Sig.
1	1096.254	8	.000

由于对微博内容的准确分析和分类需要人工介入，本次的分析只考虑了形式、环境、作者等方面的影响，而对内容的分析只考虑了微博长度。结果显示这些外围的因素对微博获得响应的影响较弱。下一步的回归分析重点是对内容变量的分析，包括类型、煽动性词汇、情绪、社会化倾向等方面。此外，除因变量响应与否之外，还将增加对因变量响应力度的分析。

5.2 微博的社会网络分析

5.2.1 研究背景

在 Web 2.0 时代，微博的信息传播从动机上来说是源于分散在网络各个角落的网民自发进行的一种信息传播活动，这不像是在 Web 1.0 时代那样的由某一个媒体机构点对面的线性或撒网型传播模式，而是人与人之间织成了一张网络，由这张网络中的某一个“节点”进行信息发布，并通过网络中节点间的传递产生信息传播。这张网络就是一张微博世界里的“社会网络”。

社会网络分析方法是社会学中的一个重要研究方法，它包含两个核心的概念：节点与线段。通过研究节点与节

点间的关系来呈现整体网络结构，并且解释社会网络如何影响人们的行为。在传统的 Web 1.0 时代，受到信息技术的限制，社会网络在信息传播中的作用和影响力还不那么大。但是在现在这个社会化媒体时代，尤其是在微博这样一个媒介上，公共信息的传播更为普遍，由此调动的节点的社会网络更为广泛，不同个体的社会网络间的相互作用也更为明显。因此，目前在以社会网络为信息传播的基本结构方面，微博表现得最为充分。用社会网络分析方法来对微博的信息传播机制进行分析也是较为合适的。

5.2.2 研究问题

在整个微博空间信息传播的过程中，有一个现象值得关注，那就是拥有极高粉丝数的加 V 用户的“跨界”发声。文体明星、企业家、学者都在公共事件中发表自己个人化的观点，他们这些在现实社会中各自领域的权威，通过微博这样一个平台似乎有可能获得“跨界”的权威——而在某种程度上，他们对这些事件的观点并不一定是站得住脚的。

本研究就将聚焦于微博“名人”用户的“跨界”发声行为，通过社会网络分析的方式搜集若干个案例，并采用多个案分析法对这一问题进行归纳。由于数据抓取技术和研究条件的限制，本文仅选取了姚晨、潘石屹、蔡学镛、艾米四个微博博主的 30 条源微博及其所产生的转发关系链，采用社会网络分析的方法对其中心度及网络中心势进行了分析，得到以下初步结果。

5.2.3 研究内容及结果

1) 中心性分析

表 23 : 姚晨的社会网中心性分析

序号	相对中心度	网络中心势	微博文本
1	91.279	91.18%	微笑是一生的财富：）
2	36.742 （任志强 57.955）	57.57%	红配绿,生活还是要继续。[偷笑]@ 任志强
3	99.630	99.63%	两位“潇洒哥”不带头盔，甩着乌黑靓丽的头发，跨着小摩托，在东四环车流中自由穿梭…哥，你们这是在拍戏吗？剧组给买保险没？
4	91.097	91.07%	谢谢你们周到的服务，尤其是戴紫色围脖的香港妹妹，对每一位乘客都投以极大的耐心，令入印象深刻。对了，空姐制服很不赖，帅气又性感噢。[花心]
5	99.890	99.89%	折翼的小天使，就这样飞走了。。。她们在哭泣，她们在问：叔叔，为什么，我们很乖，为什么，你这样的恨我们。。。

表 24 : 潘石屹的社会网中心性分析

序号	相对中心度	网络中心势	微博文本
1	99.338	99.34%	无论政治或法律，科学或工业，人类世界正在经历革命性的转变。陈旧的伦理规范、道德标准和过往的生活方式已经不再适合当今进步与发展时代的需要了。
2	99.355	99.35%	“那种认为‘灵魂会随肉体死亡而消逝’的想法，就如同想象‘笼破鸟必亡’一样，尽管鸟儿对于笼之被毁无可畏惧。肉体如笼，灵魂如鸟。笼虽破，鸟尚存，而且，其感觉更强烈，其感知更敏锐，其快乐也大增。”——《若干已答之问》
3	88.384	88.27%	困在上海机场，一会儿请任总吃饭，@ 冯仑也困在上海，一会他过来
4	33.766 （创业家杂志 58.874）	58.51%	“潘币”调侃背后是对高房价的不满 http://t.cn/asPvSi
5	52.045 99.257	98.87%	有微博上的朋友问，黑钵里装的是什么吃的？不是吃的，是松塔和干柠檬。装点着圣诞的气氛。
6	74.329	74.26%	早上我与一住宅开发商老板通电话。我：“你们降价吧，给大家一个台阶下。我给你们背黑锅，潘都成计量单位了。”他：“不敢降，降了客户就来砸售楼处了。”
7	67.192	67.13%	春兄：我从来没有当过官，瞎说，不许笑我。都回家去，在网上办公审批。每件审批向全社会公示。办公楼改成科技馆（在波士顿见过）学生们免费在里面学习、上网、讨论。
8	99.275	98.90%	村口的那棵白杨树。
9	99.661	99.66%	在山里走遇到一刺猬，不知死活。等我回来看没有走动。
10	99.693	99.69%	给大家发一张东京银座的夜景。晚安。

11	99.505	99.45%	昨天在 CCTV 录《对手》后，我给栏目组一个小建议，不要以吵架为主，要有讨论问题、理性、磋商的气氛。@ 袁岳先说我是房地产界的恐怖份子，@ 任志强上来又与 @ 袁岳对吵，@ 叶檀也不示弱。让我想起了文革。
12	99.752	99.70%	今天我再送 50 本《我用一生去寻找》的书并附带“书签”，网上网下一起交流。前 50 名留全姓名、地址和邮编的朋友们。
13	18.720 (王利芬 34.123)	33.81%	今天北京空气质量。
14	99.854	99.36%	北京网球公开赛（一）
15	83.980	83.94%	人心的改变是根本的改变。
16	0.000 (玖月贰拾伍日0.110)	0.11%	十几年前，我看一些宗教方面的书，常觉得自己快开悟了，一天我给 @ 张欣说明天我可能就开悟了。她说，总想着自己要开悟的人，离开悟还远着呢。今天回想起当初我的想法是多么的幼稚可笑。
17	89.173	89.15%	最重要的改变人心。

表 25 : 蔡学镛的社会网中心性分析

序号	相对中心度	网络中心势	微博文本
1	99.254	96.18%	创业时选择使用什么技术来实现项目，除了要考虑技术本身适合项目与否，还必须考虑有多少人会这项技术。选择小众的技术，很可能招人方面会遇到大问题。我知道几个采用 Ruby 与 Common Lisp 等技术的项目，招人上特别困难。
2	99.115	99.08%	知名程序员 Simon Peyton Jones 说不管有多么不起眼、多么老套、多么不重要，都要去做，因为这就是你开始的方式。一旦开始，就会发现计算机科学不管看来差别多大，都有相似之处——几乎每样东西都很有趣，因为这门学科一直走在你前面。它不是定型的东西。它的范围在不断地扩张，等着你去发现。
3	98.540	93.94%	商场内摆满了月饼，原来中秋快到了。小时候家里开面包店，中秋节可是我们卖月饼赚钱的重要日子。但中秋节一过，月饼顿时乏人问津。卖剩的月饼，我们只好自己吃掉，早餐月饼、午餐月饼、晚餐月饼，天天过著吃月饼的悲惨生活。
4	99.286	98.87%	去书店看到一本书《智能 Web 算法》（Algorithms for Intelligent Web），买回来正在读，觉得不错，可以推荐。许多网站后台需要推荐功能、搜索功能 ... 等算法，可以看这本书入门。
5	99.425	98.63%	如果你是手机应用开发者，你可以试图把简历做成一个 App，比方说「Meet Jerry Tsai」或者「There is something about Jerry」，放上 Marketplace 免费下载，顺便展示你的编程功力。[威武]
6	99.265	97.14%	台北市全面开放免费无线上网，昨天是第一天，不过也许是受到天候影响，许多民众纷纷反映，「Taipei Free」的讯号微弱，许多路段不是连不上，就是常常断线，没有想像中的好用。受限于 WiFi 的技术，讯号会出现死角，加上昨天天候不佳，AP 功率多少会受影响。... 不过至少是个开始啦！值得鼓励。[鼓掌]
7	95.448	95.38%	如果你是一个在校生，希望以后进入 IT 行业，但没有明确的目标以后要做什么技术领域，那么我会建议你把时间花在 JavaScript 以及相关框架上。这是目前 " 性价比 " 最高的技术。性＝用途广泛性（跨前后端与各种平台），价＝学习的代价（耗费时间与难易程度）。

表 26 : 艾米的社会网中心性分析

序号	相对中心度	网络中心势	微博文本
1	98.980	97.46%	微博，但自我感觉那是多么良好啊！瞧，我多勇敢，我多善良，我多关注温州事故啊！拜托，别自己恭维自己了，你不过就是跟个风而已，你也只会跟风。

2) 传播模式类型

通过上述统计发现，在 22 个社会网络关系中，微博博主具有较高的中心性（21 条大于 90%，1 条为 89.15%），他们的社会网络关系图一般表现为“中心发散

模式”（见图 16）。

在这种模式中，绝大多数转发关系都是直接发生于源微博上，各个转发用户之间并不存在太多关系，这种信息传播的模式主要就是依靠源微博博主极高的粉丝数量达到的，不存在其他若干个起到中介作用的节点。在这种模式下，源微博博主对该话题的影响力非常高，占据着极高的地位。

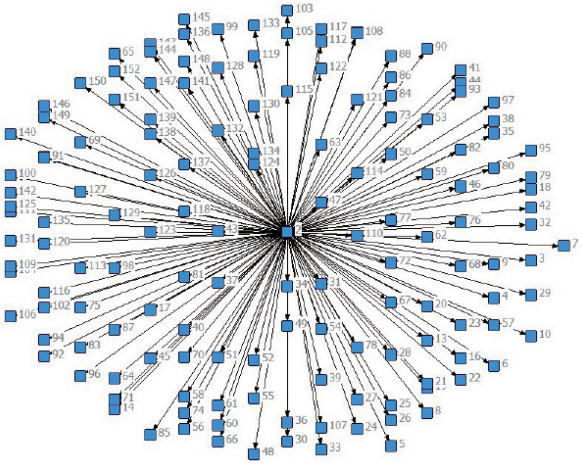

图 16 : 中心发散模式

在中心发散模式的基础上还存在一种变型，即在信息传播的大节点之外还存在若干个凝聚子群，可以称之为“二级转发模式”（见图 17）。这些凝聚子群扩大了这张转发关系网络的传播效果，有着较强的中介性。在这种模式下，这些凝聚子群的中心节点与源微博博主一起构成了信息传播结构的中心地位。

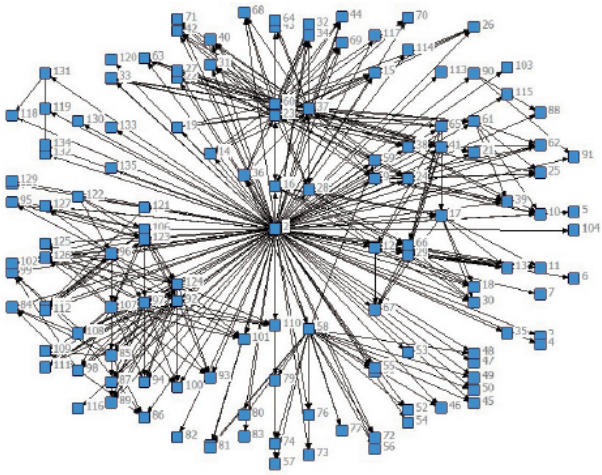

图 17 : 二级转发模式

在剩余的 8 个社会关系网络中，又主要存在两种传播机制。一类为“二度接力”模式，即源微博在发布后，经历了一次“重要转发”，承担这次转发功能的博主又作为中介引发了基于其的第二次广泛传播，但这次传播过程是几乎独立于源微博的转发关系的，潘石屹（4）就是这种类型的代表（见图 18）。在这种模式下，源微博博主和引起第二次新的传播过程的中心节点一起占据了传播的高地。

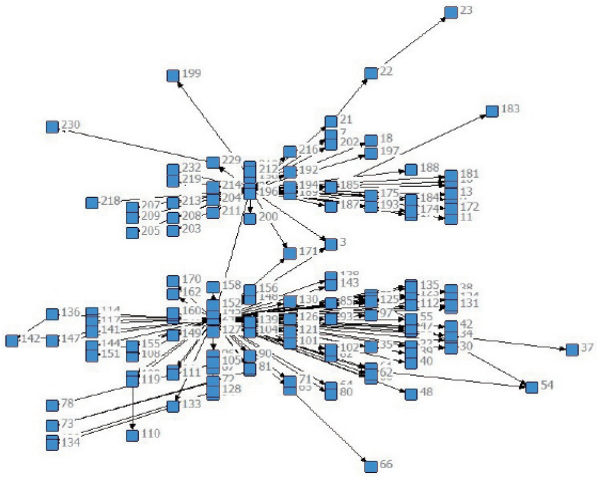

图 18 : 二度接力模式

另一类则较为少见，整个社会网络的传播具有极低的中心性,具有最高中心度的博主只有0.110的相对中心度，潘石屹（16）就是一例（见图 19）。在这种“无中心模式”下，传播的社会结构是无序的。

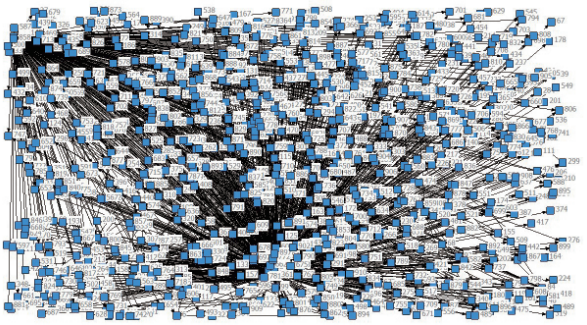

图 19 : 无中心模式

5.2.4 研究局限及下一步计划

由于所抓取的数据并非随机数据，故而抽样样本代表性有限，无法将话题内容、博主信息等变量与社会网络模式进行相关性分析，也无法对社会权力的迁移进行科学的界定和判断。同时，由于样本数量有限，所以本研究

进行的社会网络模式归纳也需要进一步研究和确认。

所以，本研究下一步打算着重在数据挖掘技术上进行探索，获取较为具有代表性的抽样样本，扩大博主选取的范围，形成 10 人左右的名人博主样本，300 ～ 400 个左右的微博样本，在这一样本数据和资料的基础上，逐渐归纳和建构出微博空间加 V 用户微博传播的若干种模式。同时将研究的关注重点放在这些加 V 用户在微博传播网络中的地位与现实中的比较，并从社会结构的角度进行分析。

6 总结和展望

综上所述，是醒客工场研究员们针对微博传播课题，进行的第一次迭代探索，意在建立研究的概念框架雏形、提出问题，征集各界的意见、反馈，同时也是为了探索醒客工场跨学科合作研究的工作模式和方法。在下一个阶段，我们将从以下几个方向进一步深化我们的研究：

- 扩大样本数据的采集规模和形式。在目前的报告版本里，我们采集的微博研究数据，无论是规模上，还是完整性、有效性上，均存在一定的限制。下一阶段，一方面，我们进一步强化微博样本数据的采集、筛选；另一方面，我们将拓展调研数据采集的形式和内容，不局限于计算机程序自动搜索、抓取的微博、作者和评论转发数据，也将通过问卷调查、用户访谈等形式收集用户偏好、使用习惯等方面的样本数据。
- 形成基本的研究结论。在满足有效的样本数据的前提下，我们将进一步完善研究方法，包括检验、调整用于规律预测的数学公式等，形成微博传播规则研究的初步结论。
- 构建课题研究的内在统一性。目前的研究框架，是在微博传播规律研究的大主题下，研究员们分别从传播半衰期、引爆点、社交关系、心理学等角度进行探索。由于只是一个初步的概念框架，上述模块之间相对独立，没有形成彼此的理论关联。随着样本数据的进一步丰富，以及分析结论的得出，我们将开始将上述研究模块进行衔接，构建醒客工场关于微博传播规则的初步理论体系。
- 开发辅助研究软件。为了更好地提炼课题研究中积累的经验、方法和技术，我们会针对研究领域中相对成熟的算法或者方法模式，研发相关的软件工具（比如微薄信息半衰期分析软件、用户情绪文本分析工具等），这不仅是为实现未来醒客工场项目研究平台化的目标做前期准备，而且我们也会酌情将其中部分技术开源发布，以方便其他研究者进行相关研究工作。
- 增加应用案例分析。随着分析结论的形成和理论体系的初步建立，我们将更多实证案例进行考察，例如商品的流行度、社会热点事件等，以此探索和实践我们研究微博的初衷——通过数据来呈现一个全新的认识世界的视角。

主要参考文献

[1] Kwak, H., et al., What is Twitter, a Social Network or a News Media? [A], in WWW 2010, 2010: Raleigh, North Carolina, USA.

[2] Leskovec, J., L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. [A] in the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009. Paris, France.

[3] Ratkiewicz, J., et al. Detecting and Tracking Political Abuse in Social Media. [A] in International AAAI Conference on Weblogs and Social Media. 2011.

[4] Java, A., et al., Why we twitter: understanding microblogging usage and communities, [A] in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis2007, ACM: San Jose, California. p. 56-65.

[5] Kivran-Swaine, F., P. Govindan, and M. Naaman. The Impact of Network Structure on Breaking Ties in Online Social Networks: Unfollowing on Twitter. [A] in 28th international conference on Human factors in computing systems. 2011. Atlanta, Georgia, USA.

[6] Kwak, H., H. Chun, and S. Moon. Fragile Online Relationship: A First Look at Unfollow Dynamics in Twitter. [A] in 28th international conference on Human factors in computing systems. 2011. Atlanta, Georgia, USA.

[7] Westman, S. and L. Freund, Information Interaction in 140 Characters or Less: Genres on Twitter, [A] in IIiX 2010,

2010.

[8] 夏雨禾 . 微博互动的结构与机制——基于对新浪微博的实证研究 [J]. 新闻与传播研究，2010（04）：60-69.

[9] 喻国明，欧亚，张佰明，王斌 . 微博：从嵌套性机制到盈利模式 [J]. 青年记者，2010（21）：18-21.

[10] 张佰明 . 嵌套性：网络微博发展的根本逻辑 [J]. 国际新闻界，2010（06）：81-85.

[11] 袁毅 . 微博客信息传播结构、路径及其影响因素分析 [J]. 图书情报工作，2011（12）：26-30.

[12] 赵云泽，付冰清 . 当下中国网络话语权的社会阶层结构分析 [J]. 国际新闻界，2010（05）：63-70.

[13] 李莹 . 微博对日本地震相关信息传播的正负效应 [J]. 现代传播，2011（10）：163-164.

[14] 高承实，荣星，陈越 . 微博舆情监测指标体系研究 [J]. 情报杂志，2011（09）：66-70.

[15] 王晓光 . 微博客用户行为特征与关系特征实证分析——以“新浪微博”为例 [J]. 图书情报工作，2010（14）：66-70.

[16] 李多，周蔓仪，杨奕 . 网络平台对于政府与网民之间关系建设作用的探索——以伍皓的微博为例 [J]. 新闻知识，2010（08）：49-51.

[17] 杨晓茹 . 传播学视域中的微博研究 [J]. 当代传播，2010（02）：73-74.

[18] 王建磊 . 社交型媒体与变形的新闻 [J]. 新闻记者，2010（09）：65-69.

[19] 余伟利 . 从博客到微博：网络问政两会的媒体应对 [J]. 现代传播，2010（06）：143-144.

[20] 蔡骐 . 粉丝型受众探析 [J]. 新闻与传播研究，2011（02）：33-41.

[21] 闵慧泉 . 真实与虚拟：新媒介环境下的追问 [J]. 现代传播，2010（02）：110-113.

[22] 张月萍 . 微博客对网络新闻评论的影响 [J]. 新闻大学，2010（03）：118-119.

[23] 邱均平 . 信息计量学 [M]. 武汉：武汉大学出版社，2007：615.

[24] 杜子建 . 微力无边 [M]. 沈阳：北方联合出版传媒（集团）股份有限公司，2011：125-127.

[25] 喻国明 . 微博：一种新传播形态的考察 [M]. 北京：人民日报出版社，2011：13-15.

[26] PageL，BrinS，MotwaniRetal.Thepagerank citation ranking:Bringing order to the web[R].Stanford Digital Libraries，1999.

[27] Prasad, J. A comparative study of rumours and reports in earthquakes[J].British Journal of Psychology, 1950, 41(3-4): 129-144.

[28] 刘志明，刘鲁 . 面向突发事件的群体情绪监控预警 [J]. 系统工程，2010，28(7): 66-72.