基于多重关系的语义地图算法

——探索语言系统中的隐性控制规律

陈振宁　陈振宇
（文华学院，湖北武汉，430074；复旦大学，上海，200433）

摘　要：语义地图研究语言中概念个体间的蕴涵关系，即用“图”来研究概念个体通过共现交际关系得到的控制路径。语义地图的理想绘制方法是一种“关系的关系”即多重关系，多元共现通过独立的两两共现来确定。对理想绘制方法在概率上进行一般化，得到三种“赢家”算法，对独立两两共现频次高的边做出不同倾斜，从而具有更好的隐性控制规律概括性。其中赢多输少算法对处于控制边缘的“输家”边有更清晰地分层作用。

关键词：隐性规律，图论，语义地图，算法，多重关系

一　基本理论

1.1 语义地图

语义地图（Semantic map）理论最初的目的是为了研究跨语言的互相有“功能”联系的多个形式的蕴含共性。将多个形式所表达的有联系的功能绘制为“图（graph）”，其中：

图的节点（nodes）：按“功能”或“意义”划分的“基元（analytical primitives）”（Cysouw，Michael，2007），即由研究者通过对跨语言的某类语言形式进行比较，最后得到的语义成语用的基本概念单位，如 Cysouw（2003）对人称的基元分解如表 1：

基元符号	基元说明	对应的人称形式（汉语）
1	说者	“我”
2	听者	“你”
3	第三方	“他/她/它”
12	说+听	“咱们”、“我们$_1$”（我们两个人去上课吧）
123	说+听+三	“我们$_2$”（我们几个人一起去上课吧）
13	说+三	“我们$_3$”（我们去上课了，你呢？）
23	听+三	“你们”
33	多个第三方	“他们”

表 1　Cysouw（2003）对人称的基元分解

图的边（edges）：基元共现的情况，即如果调查发现若干基元共现于同一形式中，则认为这几个基元间有关系存在。例如，Cysouw（2003）调查的数据中，3 和 33 在跨语言的 125 个形式中共现过，那么 3 和 33 两个基本概念之间是有关系的，在这两个节点间用边连接以表示这种关系。

但是，基元个数更多添加边成为一个问题。以 3 个节点构成的三元图中为例，全连通的 4 种子图都能实现“三个基元共现于同一语言形式”的关系，如图 1：

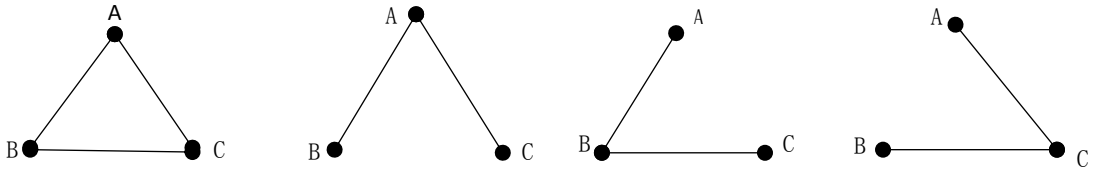


图 1　全连通三元图的 4 种子图

Haspelmath（2003）、Haan,Ferdinand（2004）提供的绘制方法：三元图的关系由二元图

两两共现的情况决定。如表 2 和图 2：

	A	B	C
L1	X1	X2	X3
L2	X4	X4	X5
L3	X6	X7	X7
L4	X8	X8	X8

L=语言；X=形式；A、B、C=功能

表 2　三元共现关系

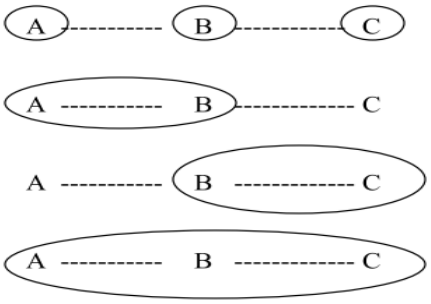


图 2　理想的三元图

理想情况下，ABC 三者共现时，，如果只有 AB 共现和 BC 共现，那三者关系自然是 A-B-C。相应的，若是只有 AC 共现和 BC 共现，那三者关系是 A-C-B。其余可以此类推。

但是，现实调查的数据往往是不理想的。即 ABC 三者共现时，同时又有 AB、AC、BC 共现，于是形成一个 A-B-C-A 的回路，即“空地图”，如图 4。

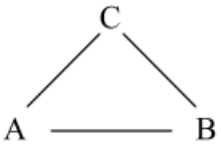


图 4　空地图

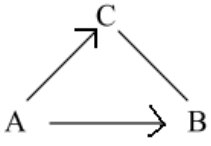


图 5　有向语义地图

为了使空地图有意义，解决方法之一是通过概念的历时演变关系，给边加上方向，如图 5。本质上将无向图变成有向图，有向图的连通性是单向的，就断开了回路，保持了蕴涵关系。

这个解决方法的问题在于：必须明确历时关系，否则无法操作。但在语言调查中，历时关系往往难以确定。实际上，研究者往往就是希望从调查的共时数据中对历时进行推测。

Cysouw（2007：19）注意到，基元共现的情况其实是“不均匀”的，即不同共现的频次有很大差异。他给出了人称 8 个基元共现的频次表，如表 3：

频率	1	2	3	12	123	13	23	33	yes 数
125			+					+	2
97				+	+				2
84		+					+		2
29	+					+			2
17							+	+	2
10	+		+						2
7		+	+						2
3					+	+			2
3	+	+							2
2				+		+			2
2						+	+		2
2			+				+		2
1						+		+	2
1	+			+					2

频率	1	2	3	12	123	13	23	33	yes 数
3	+	+	+						3
2				+	+		+		3
1					+	+	+		3
1				+	+			+	3
1	+			+	+				3
35	+			+	+	+			4
18				+	+	+	+		4
11				+	+	+		+	4
6		+	+				+	+	4
5		+		+	+	+			4
4		+		+	+		+		4
1			+	+	+			+	4
5				+	+	+	+	+	5
2	+			+	+	+	+		5

1	+						+		2
100				+	+	+			3
5			+			+		+	3
4		+				+	+		3

1	+	+		+	+	+	+		6
1		+		+	+	+	+	+	6
1	+	+		+	+	+	+	+	7

表 3　人称 8 个基元的共现情况表

不同共现的频次差异说明：在被调查的人类语言中，基元之间的不同关系，有着不同的权重。其中有些关系的权重高，是“主流”关系，有些关系的权重低，是“非主流”关系。在主流关系中占中心地位的基元，往往是这组基元集合的“控制中心”。

也就是说，我们要用“交际和控制”的视角来审视相关问题。

1.2 交际与控制

交际与控制的概念来自对社会网络的研究。人类社会是通过“行动者”之间不停的交际形成的系统。其中，行动者可以是单个的自然人个体，也可以是自然人组成的各种组织个体。行动者之间一旦发生交际，就产生了关系。各种关系交织起来，我们的社会就成为了一个交际的网络，或者说交际的图。行动者是图中的节点，各种关系是图的边。数学中的“图论（Graph Theory）”正是从各种图研究中抽象出来的数学理论。

推广来说，任何具有个体离散性的复杂系统都可以看做个体和交际关系的结合而成的图，作为人类社会一个部分系统的语言系统，是“单位的组合/聚合系统”（索绪尔）。其中 “单位”即是图的节点，单位之间的各种“组合/聚合（交际）关系”是图的边。不同单位在系统中的“价值”就是由其交际关系决定的，同时系统的规律也体现在交际关系的网络布局中。

不同行动者参与交际的程度是不一样的，图中交际密度大的“区域”，其中的关系和行动者成为“控制中心”，对社会的“规律、制度、风俗、走向、潮流”等起着主要的甚至是决定性的作用。从抽象的图论的观点来看：

最典型的“控制中心”是辐射状星型图的中心，如图 6。其中 A 点作为辐射中心，其他所有点都只和 A 点直接交际，因此 A 拥有绝对的控制地位。

最典型的“控制路径”则是树（tree）图，如图 7。其中各点的控制层级分明，上级控制了他的直属下级，没有控制“回路”。树的特例是“链”，即所有节点相继首尾相连且没有回路，图 2 所表达的蕴涵共性就是一个链。

最典型的“无控制”是团（clique）。团是指三元或三元以上的完全图（complete graph），其中任意两点之间都可以“直接、均匀”的相连。图 4 所示“空地图”就是三元的团。因为任意两点都可以无差别相连，所以反而完全没有任何控制关系了，因此无疑语言学家认为这个“空”地图无法进行研究蕴涵共性。

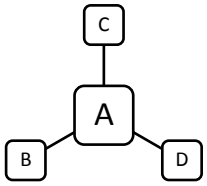


图 6　星型图

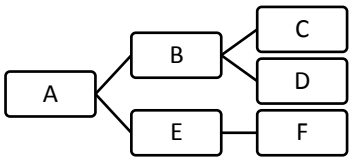


图 7　树图

1.3 显性控制与隐性控制

用图的形式来研究的控制是隐性控制，而与隐性控制相对应的是显性控制。

1、显性控制：一旦系统成员产生明确的关于某种运行规则的认识时，这一规则便“外化”于该社会，成为一条显性的机制，从而得以在纷繁复杂的社会事件中保持相对独立、静止的面貌，并反过头来要求社会必须遵从于它。社会关系中来说，就是被加冕为王，有了王冠和“大义名分”，从而用法律和最大的权力来要求所有人遵循其守则。语言系统中，明确的“形式标记”就是显性控制的体现，标记必须出现就是外化的显性规律。

2、隐性控制：在显性化之前，一个未曾事先规定任何规则的社会，在其自身的运行中仍然会自发地形成一套运行机制，但它仅仅是现在的、当下的、自动地形成着的，每一个人都卷入其中，而未必知道它，“不识庐山真面目，只缘身在此山中”。社会关系中，其实任何一个社团都存在这样的隐性中心，如一个朋友圈里的核心成员和他的亲密圈子，如国王身边若干个最具有影响能力的大臣集团。甚至这个隐性中心未必是“强势”的，如通常情况下，一个家庭有了小孩以后，恰恰小孩极弱极需要依赖旁人照顾，反而使得父母的“二人空间”大大减少而一切家庭活动多是围着小孩在转，从而小孩成为了自然的“家庭中心”。

语言系统中，凡是没有被“形式标记”明确化的一切控制都是隐性控制。同样的，隐性中心不一定“强势”，实际上语言系统中居于句法上位的功能性成分往往显得“弱势”，但正是因其“依赖性强独立性弱甚至极弱”，功能性成分得以联系独立性相对强的实词成为句法中心。

类型学研究的“蕴涵共性”本质上就是图所展示的一种隐性控制。类型学研究的是跨语言的控制能力，在跨语言中形式不定，此语言的这几个概念有明确形式标记，到彼语言换成了相关另外的概念有明确形式标记，如同各自在自家地盘都是王，但诸王碰头，就只有“拉在一起比一比”才知道谁是“王中王”了。

隐性控制的“隐蔽”性，无法直接定性描写，只有用定量来做研究。但是隐性的内部控制很难把握，它具有模糊性、即时性、变化性等特征，而且亲身经历它的人也往往缺乏共识。所以对隐性控制的探索至少要衡量两点之间的平衡：

1、概括充分：隐性控制难以识别，调查数据一旦纷繁复杂而算法过于简单，可能根本概括不出什么规律来，因此需要有能够进行概括的算法。

2、概括适度：同时，隐性控制又总是处在形成的过程中，如果控制机制本身不强，过强的识别算法反而会坏事。“过犹不及”，又称为“过度理解”（over-understanding），它与“错误理解”（mis-understanding）一样，都是对事实的非真实的反映。

也就是说，在绘制语义地图的时候，如果“不得不”有回路，那就还是得有回路。

二　多种算法

2.1 已有算法——完全加权算法和完全关联算法

Cysouw（2007：19）提出了将基元共现频率累加为各边权重的基本思想，绘制出加权的人称语义地图。他的加权算法非常简单：

1、如果基元 a、b 两两共现，就将两个基元之间连上一条边 a-b，且为这条边加上共现频次 f_{ab} 为权重为；

2、如果基元 a、b、c 三者共现，就认为它们两两之间全部存在同一关系，于是两两之间全部连上边，一共是 3 条边 a-b、b-c、a-c，同时为 3 条边都加上 f_{abc} 为权重；

3、以此类推，如果 n 个基元共现，就认为它们两两之间全部存在同一关系，于是连接上所有 n*(n-1)条边，同时为所有 n*(n-1)条边都加上 f_n 为权重。

简言之，该算法直接把语义地图处理为了任意两点之间都有直接联系的完全图，Cysouw 用完全图加权算法得到表 3 数据生成的语义地图，其权重矩阵如图 8。

	2	3	12	123	13	23	33
1	8	13	41	40	68	5	1
2		16	12	12	4	101	8
3			1	1	5	8	137
12				286	181	34	20
123					184	35	20
13						35	24
23							30

图 8　完全图形式的人称语义地图权重矩阵

但这恰恰和蕴含共性研究的基本思想是相反的。如 1.2 所述，完全图就是空地图，完全图加权的算法，就会得到和语义地图绘制要求相悖的结果，如例 1。

例 1：人称 3、13、33 的局部语义图。

在 Cysouw 的调查数据中（表 3），人称基元 3、13、33 三者共现且频次为 5。于是从图 8 我们可以看出，完全图加权算法把 3-13、3-33 和 13-33 都连接起来，并且都加上权重 5[1]。

但是，如果考察这三者的独立的两两共现的情况，我们会发现：

3	**13**	**33**	**独立的两两共现频次**
+	+	+	5
+	**+**		**0**
+		+	132
	+	+	19

表 4：3、13、23 的共现情况

这恰恰是绘制语义地图的理想情况，因为 3 和 13 从不独立两两共现，所以完全可以把 3-13 边彻底删除，从而只保留 3-33-13 这条链。但完全加权算法的机制下，从不独立两两共现的 3-13 依旧是连接在一起的，并有权重 5。

亦即，完全加权方法并**不兼容**“理想的绘制方法”。

另外，郭锐（2012：115-116）引入了“关联度”的概念，两个概念间关联度的计算公式为：

$$A = S1\ S2 / (W1 + W2 - S1\ S2) \times 100$$

其中，“S1 S2”指基元 1 和基元 2 在数据中“任意”共现的频次，W1 指具有义项 1 的总频次，W2 指具有义项 2 的总频次。简单地说，就是将完全加权按一定比例缩放，也是把每一个记录局部视作完全图，计算的是一种“**完全**关联度”。

无疑，完全关联算法也**不兼容**理想的绘制方法。

为此，本文考虑设计能够兼容理想洽制方法的算法。

2.2“赢家通吃”算法

本算法的基本操作思路是：赢家通吃，输家不吃。具体如下。

对调查数据中的一个记录里，如果有 n(n≥3)个基元共现：

1、两两结对，形成 n*(n-1)个对子；

2、查其他所有记录，计算每个对子的独立共现频次；

3、按两两独立共现频率在对子之间竞争；

4、前 n-1 个频率大的对子成为“赢家”，其余“输家”；

5、赢家连接，都获得本记录的 100%加权；

6、输家什么都没有，不连接，不加权。

这样计算下来，每个记录构成的局部语义图就是一颗控制关系明确的树，如例 2：

例 2：人称 12、123、13 的局部语义网。

表 3 中人称 12、123、13 有一条三者共现的记录，频次高达 100。3 点能形成 3 个对子：12-123、123-13 和 12-13，我们根据调查数据计算这三个对子的“独立两两共现频次”。

独立两两共现频次由两个分量组成，我们将其中和这三者共现有关的共现情况都截取出来，得到表 5：

[1] 13-23 因还在其他共现中出现，因此最后权重为 24。

序号	频次	1	2	3	**12**	**123**	**13**	23	33	yes 数
1	97				+	+				2
2	3					+	+			2
3	2				+		+			2
4	**100**				+	+	+			**3**
5	2				+	+		+		3
6	1					+	+	+		3
7	1				+	+			+	3
8	1	+			+	+				3
9	35	+			+	+	+			4
10	18				+	+	+	+		4
11	11				+	+	+		+	4
12	5		+		+	+	+			4
13	4		+		+	+		+		4
14	1			+	+	+			+	4
15	5				+	+	+	+	+	5
16	2	+			+	+	+	+		5
17	1	+	+		+	+	+	+		6
18	1		+		+	+	+	+	+	6
19	1	+	+		+	+	+	+	+	7

表 5　人称 12、123、13 的共现情况

表 5 中第 4 行就是我们要处理加权的局部记录。

独立两两共现频次的第一个分量，是排他的两点共现，如表 5 的 1-3 行。

独立两两共现频次的第二个分量，是其他多点共现中，不包含本记录所有 3 点但包含其中两点的共现，如表 5 的 5、6、7、8、13、14 行。

而表 5 中 9、10、11、12、15、16、17、18、19 行，包含了本纪录全部 3 点，对分辨本记录无作用，就不计算在内。

由此，计算三个对子的独立共现情况，如表 6：

12	**123**	**13**	**独立共现频次**	**竞争结果**
+	+		106	赢家+100
	+	+	4	赢家+100
+		+	2	输家+0

表 6　12、123、13 的独立两两共现频次

其中，12-123 单独共现了 97 次，包含 12-123 但不包含全部三点的记录频次之和为 9 次，因此 f12-123=97+9=106；

123-13 单独共现了 3 次，包含 123-13 但不包含全部三点的记录频次为 1 次，因此 f123-13=3+1=4；

12-13 单独共现了 2 次，包含 12-13 但不包含全部三点的记录没有，其频次为 0 次，因此 f12-13=2+0=2。

可见，12-123、123-13 的独立共现频次 106 和 4 大，是赢家；12-13 的独立共现频次 2 小，是输家。

赢家通吃 12-123、123-13 全部加权 100；输家不吃 12-13 无加权，结果如图 9。

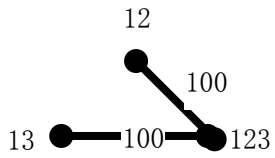


图 9　赢家通吃算法得到的 12、123、13 局部语义图

无疑，在理想情况下，那些两两独立共现频次为 0 的对子会成为“彻底的绝对输家”，其对应的边就会得到 0 权重，从而不连接。也就是说，赢家通吃算法在理想情况下就是理想的绘制方法。

换言之，我们确实设计了一个**兼容**理想绘制方法的加权算法。

2.3“赢家多吃”算法和“赢多输少”算法

再回头来审视例 2 中的人称概念 12、123、13 的共现，我们注意到一点：

赢家中 12-123 的独立共现频次确实很高，作为“大赢家”获得全部加权没多大问题，但同样是赢家的 13-123 的独立共现频次就陡然降到了 4，与“大赢家”相比，实在是一个太弱的“小赢家”。但赢家通吃算法不考虑这种“赢多赢少”的差异，都给予 100%加权，就带来一个节 1.3 提到的问题：

会不会因此造成小赢家相对大赢家被“过度概括”呢？

更进一步看，12-13 只有 2 次独立共现，作为输家是没错，但是还不能算是独立两两共现频次为 0 的彻底输家，就直接和彻底输家一样得到 0 权重。同时，12-13 的 2 次输和 123-13 的 4 次赢对比，输得也不算“很惨”，赢家通吃却造成了“小胜小输”之间的“天壤之别”，那么还是节 1.3 提到的问题：

会不会因此造成赢家相对输家被“过度概括”呢？

由此本文设计了另外两个算法：赢家多吃和赢多输少。

赢家多吃的基本原理是：多赢多吃，少赢少吃，输家不吃。

赢多数少的基本原理是：多赢多吃，少赢少吃，输家少少吃。

即是在赢家通吃的竞争机制之上，修改一下最后的加权分配：

1、“赢家多吃”的比例分配

计算所有两两对子的独立共现频次之和 sum（$f_{独}$）；

赢家对子连接，按比例分配加权，即每个赢家对子得到的加权为：$F_{独i}$/sum($f_{独}$)；

输家对子还是不连接，加权为 0。

对例 2：大赢家 12-123 得到的加权为 106/（106+4+2）*100≈94.6；小赢家 123-13 的加权为 4/(106+4+2)*100≈3.6；输家 12-13 的加权为 0，其结果如图 10-1。

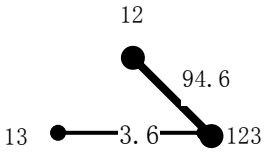
2、“赢多输少”的比例分配

计算所有两两对子的独立共现频次之和 sum（$f_{独}$）；

赢家对子和输家对子都可以连接，按比例分配加权，即每个赢家对子得到的加权为：

$f_{独i}$/sum（$f_{独}$）；

对例 2：大赢家 12-123 得到的加权为 106/（106+4+2）*100≈94.6；小赢家 123-13 的加权为 4/(106+4+2)*100≈3.6；输家 12-13 的加权为 2/(106+4+2)*100≈1.8，其结果如图 10-2。



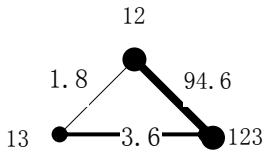


图 10-1　赢家多吃算法的例 2 结果　图 10-2　赢多输少算法的例 2 结果

同样，在理想的情况下，只要输家对子是独立两两共现频次为 0 的“绝对输家”，其权重还是 0。

所以赢家多吃和赢多输少算法还是**兼容**理想绘制方法的。

2.4 算法评估

本文设计的算法，加上 2.1 所述已有的两种算法，一共 5 种算法。编制程序根据表 4 提供的人称概念调查数据绘制语义图，如图 11、图 12 和图 13。

- 完全加权和完全关联算法

完全加权算法和完全关联算法生成的语义图相似，后者基本上是前者的比例缩放版。

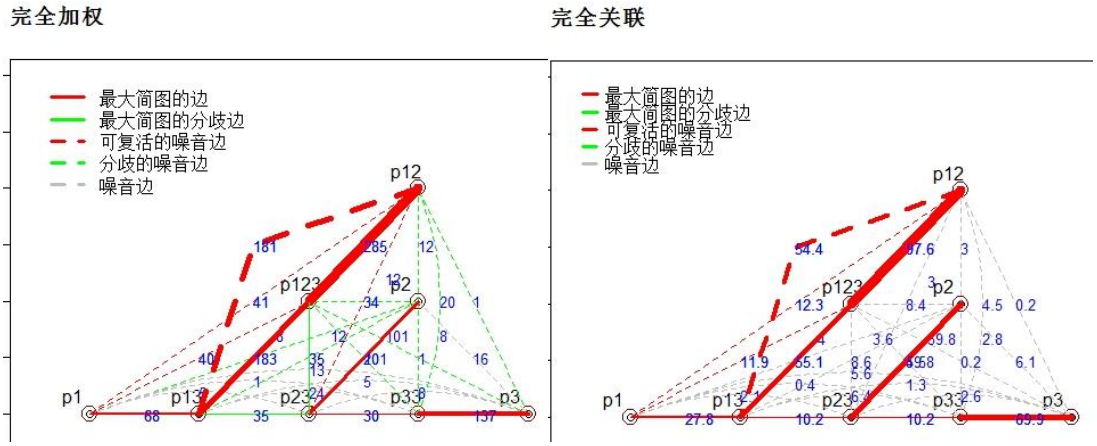


图 11　完全加权和完全关联算法得到的人称图

因为这两种算法认为表 4 中每一行记录都是互相无关的，将所有基元任意两两连接，所以最后生成的全图有最大的边数，也就是说，在连通度上来说成为了一个“完全图”。

当然，毕竟表 4 各行记录之间的频次有差异，所以这个完全图各边的权重还是有差异的。不过，当我们为了提取“控制路径”而强行删除任何可能形成的回路的时候，就会遇到一个问题。

图 11 中红线就是删除所有回路后得到的最大树图，本文称之为最大简图。Cysouw 根据自己的“语言学知识判断”删除回路后得到的简图基本和图 11-1 红色实线部分一样，但很明显：大量的被删除的边权重都很“重”，甚至比保留的边权重还重得多，图 11-1 中的红色虚线就是这样的边。

这就让人疑惑：这些权重这么重的边应该删除么？是不是应该让它们“重新获得被保留的资格”呢？因为如果回路实在无法被删除的时候，我们似乎也不应该“强删”。因此本文将这些红色虚线称之为“可复活边”。

这就又引发了一个问题：其实，可复活边中，有些边在理想绘制方法中是根本不会存在的。如图 11-1 中的 1-123，完全加权的权重高达 40，完全关联的关联度也有 11.9，各自在其全图里都是比较重的。但考察表 4 我们会发现，1-123 的独立两两共现频次为 0。

于是，1、123 两个从来没有独立共现过的点，却比其他一些独立共现过的对子还“重”。作为可复活边来看，1-123 完全没有需要复活。与之相关的回路也不是“不得不保留”的回路。

无疑，这正是因为“完全”二法不兼容理想绘制方法的原因。

这样得到的全图倾向于“均质、繁复”，要从中再抽取概括性是很困难的，对找到控制关系的目的来说，倾向于“**概括不足**”。

- 赢家通吃和赢家多吃算法

赢家通吃和赢家多吃算法是尽可能倾向所有赢家而抹杀输家的算法，生成的语义图无疑有着几个算法里最大的“不匀质”。虽然因为记录太多，这两种算法最后生成的全图总免不了有些边在生成是无法“直接杀掉”，但要得到最大简图还是很容易的。因为勉强“活”下来的一些局部输家边的权重不会太重。

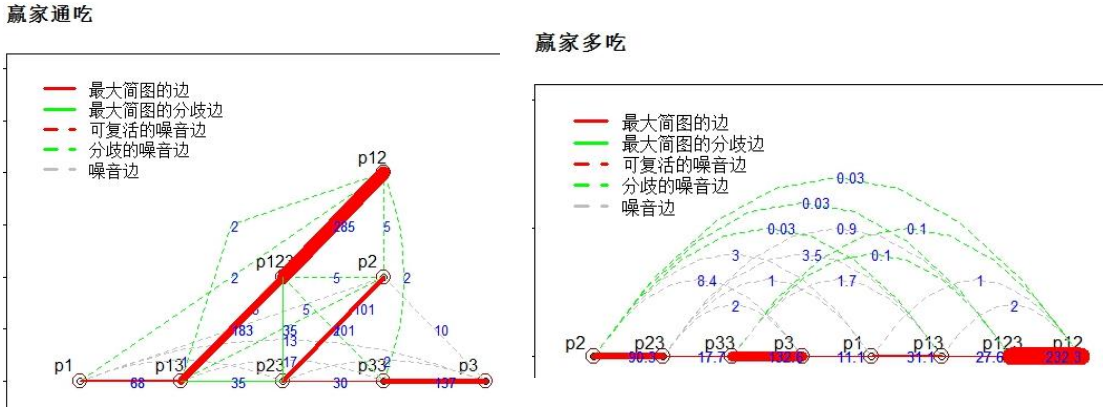


图 12　赢家通吃和赢家多次算法得到的人称图

不过，在赢家内部，它们有着截然不同的策略：赢家通吃“无差别”对待所有赢家，因此一定程度上抹杀了大赢家和小赢家的差异，使得赢家内部更倾向“匀质”；赢家多吃区分了大小赢家，在赢家内部更倾向于“差别”，但因此可能小赢家和输家之间则比前者更“匀质”些。

这一差异导致了赢家通吃和赢家多吃生成的最大简图的差异，如图 11-3 和图 11-4。

仔细看，图 11-3 和图 11-4 的主要区别在于，图 11-3 中 23（“你们”）连接着 13 或 123（“我们”），图 11-4 中确是 1（“我”）连接着 3（“他/她/它”）。

查表 4 可知，23-13 或 23-123 之间，在多次记录中虽然是赢家但都是小赢家，不过因为相关记录较多本身频次较大，每次记录得到完全加权的小赢家最后的权重还是比较大的。

反之，1-3 虽然分布的记录不多，但每次记录都赢得较多，因此在赢家按比例分配时得以“浮现”出来。

当然，因为算法抹杀了输家，所以对赢家通吃和赢家多吃算法来说，有一点是一样的，那就是最大简图外“无可复活边”。被删除的边总是权重很小“无足轻重”的“噪音”。

这样又会引发一个疑问：输家是在局部处理时被直接删除的，但局部处理时我们知道有些边输得并不多，是不是所有输家都“无足轻重”？完全不存在“必须保留的回路”、总是能得到最高概括性的赢家通吃和赢家多吃算法，会不会有“**过度概括**”的问题？

- 赢多输少算法

赢多输少算法本质上是上述 4 种算法的一个折中，对每个记录进行处理的时候，它没有绝对的无倾向和有倾向，而是根据比例关系在输赢之间建立了一个“倾斜度”，从而赢家固然有优势，输家也不至于被完全抹杀。

因此赢多输少算法和上述 4 种算法各有相似之处。

它的最大简图和赢家多吃一致，都浮现出了 1-13 这条赢的比例较大的赢家，同时压抑了 23-13 或 23-123 这对赢得比例很小的小赢家。

在人称图中它的大多数被删除边和赢家通吃、赢家多吃一样，都是小权重边，基本不可复活，在“找主流”的时候被删除也是合理的。

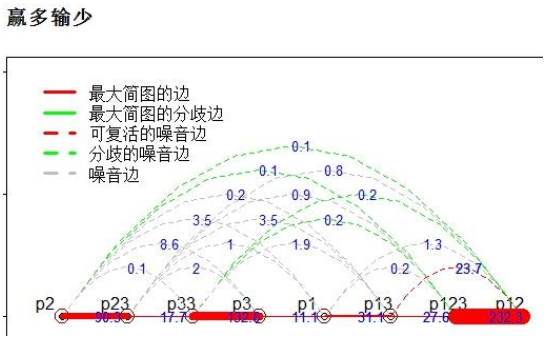


图 13　赢多输少算法得到的人称图

但是，和完全加权、完全关联一样，赢多输少没有绝对抹杀输家，就有可能导致输家“复活”。图 13 中有一条被删除边 12-13 权重达 23.7，比保留的 1-3（11.1）更高。究其原因，12-13 虽然是输家，但它们所在的共现记录频次很高。表 4 中不仅 12、123、13 三点共现频次高达 100，还有很多记录都蕴涵了 12、123、13 共现，因此作为输家的 12-13 只要不被绝对抹杀，哪怕次次皆输，最终累计的加权还是较高的。

当然，正如前文 1.1 所言，当回路不可避免时，我们似乎确实不应该强行删除。

和完全二算法不同的是，赢多输少的可复活边所在的“层次”更清晰，即它肯定是某个输得不太多的“大输家”，不可能是“绝对输家”或“输得很惨的小输家”。

同时绝对输家也可一眼即明，自然是赢多输少算法里那些两点间“不存在”的边。

被删除的各输家之间的轻重等级也是和它们的权重高低直接匹配的。

从这一点来说，赢多输少算法其实不仅仅是一个“找控制、辨主流”的算法，同时也是一个可以更好辨析“非主流、小规律”层次的算法。

当然，因为折中主义，在概括性上赢多输少不如赢家通吃和赢家多吃，如果任务仅仅是找主流控制规律，赢多输少算法反而“多此一举”了。

最后总结 5 种算法的特点如表 7：

算法	兼容理想绘制方法	概括性	过度性
完全加权		低	无
完全关联		低	无
赢家通吃	√	高	高
赢家多吃	√	高	高
赢多输少	√	中等	中等

表 7　5 种算法特性

三　赢家系列算法的理论意义

不论何种生成算法，本质上只是一种技术手段。本文不满足于完全二算法，转而研究赢家三算法，是因为完全二算法和 1.1 所述的语义地图理想绘制方法无法兼容。理想绘制方法所强调的概括原则来自“语言学家的直觉”，这个直觉就是：

首先，所有共现记录不是“一张图”，或者说单模图/单重关系，而是“多张图”，即多模图/多重关系。如表 4 里每一行共现记录是一个小图。而语义图的生成算法，本质上来说就是将多个共现记录代表的多张小图“拼”成一张大图。

第二，完全图加权算法，其实就是把多张小图视作是彼此独立，比如说其中某个三三共现和相同基元的两两共现都是无关的。因此三三共现内部是非常均匀的，那就是全都连上边并按同样的比例加权。

第三，类似语义地图等探求基元共现中的控制关系的思考，一开始就通过语言学家的直

觉，认为小图之间是“纠缠不清”的多模，因此三三共现的一个小图，还被相同基元两两共现的情况所左右。

如图 2 所示的语义地图绘制方法，其实就是把 3 个基元 A、B、C 的四幅小图叠在一起，A-B 和 B-C 放在下面当“参考”，最上面的 A-B-C 就是“描着”下面两点共现的边给“描”出来的。

只不过最开始语义地图只考虑最理想的情况。

第四，调查数据没有这么理想，于是赢家三算法就是理想状态下语义地图绘制法在统计概率上的更一般性地扩展。将原来类似的“三三共现中关系的‘有无’由独立两两共现关系的‘有无’决定”，扩展到更一般性地“三三共现中的关系的‘有无概率’由独立两两共现关系的‘有无概率’决定”。以此类推，还可以推论到更多元的情况。

理想状态下绘图法只管照着叠在下面的层层参考图“描”边之有无，可是参考图叠多了发现到处都是边，于是赢家三算法便是照着参考图的“粗细”来“描”粗细。

另外，从更一般性的意义来看，赢家多吃算法基于以下社会关系的认识：

1、最明确的关系就是排他的两两独立交往的关系，而多人关系往往是不明确的。如调查一个公司的人群，“都是这个公司的人”对调查的用处不大，稍微大一点儿的公司所谓的“同事”间可能根本不认识，和陌生人没两样。只有根据“朋友、直接上下级、工作有直接交接关系、经常性地就这两个人一起出去吃饭”这些可以排他统计的数据来对比着得到多人人群内的关系情况。

2、在“不同”的多人活动场合反复出现的两人，也更可能具有较紧密的关系。如某两人，只是在部门聚餐时才一起吃饭，那么部门聚餐的次数即使很多，也难以说明他们的关系紧密。可是如果他们在不同部门的聚餐都出现过，其他人变了而他俩一直不变，就更容易“觉得”他们肯定有某种关系。

从更抽象的层次看，任何一个系统的“个体（节点）”都不是孤立的，个体之间有“关系（边）”，这就产生了“图”的分析方法。但除了个体有关系，更复杂的情况下，“关系”之间也有关系，连接个体的关系是函数，同时又可能充当变量组成更高阶的函数，这就是多重关系的本质。

理想的绘制方法和在概率上对其扩展的赢家三算法，就是对“关系的关系”这一更高阶函数的直觉和探索。

四　余论

本文致力于开发一种有用的技术手段。技术可以极为有力地推动研究的进步。随着信息技术的发展，今天的语言学研究与 20 年、10 年前的研究，已经进入了大数据、大集成的时代，而隐藏在纷繁的语言现象背后的东西，逐渐显现出来。研究的重点开始向大数据的调查与分析转移。

但是，反过来，技术不是万能的。技术是从学者理论“直觉”上抽象出来的数学模型，本文研究赢家诸算法的动力，如前文反复强调的，正是基于学者的理论直觉，在这个直觉基础上对理想情况进行概率上的一般化。

事实上，本文对完全二算法的不满意，本来也来自使用完全二算法的前辈学者敏锐的理论直觉，因为他们在使用完全二算法时抛弃了权重所示进行简化（即理论概括）。

从语义地图中研究中引出交际与控制的理论，还有通过图对隐性控制进行研究的手段，应该能够扩展到更多研究之中：

从汉语研究来说，汉语正是公认显性形式较少而隐性规律较多的语言。

从世界语言的研究来看，形式再发达的语言，形式也只是其整个语言系统上的冰山一角，形式之下仍旧是大量语义、语用的隐性规律在运作。正如人类社会，“加冕”的各种名义统

治者和明确的法律不是万能的，绝大多数情况下仍旧是各种“圈子”里的隐性规律在运作。

参考文献

郭　锐（1993）汉语动词的过程结构，《中国语文》1993 第 6 期。

郭　锐（2012）概念空间和语义地图：语言变异和演变的限制和路径，《对外汉语研究》（第八期），北京：商务印书馆，96-130 页。

金立鑫（2014）语序类型学与普通话语序类型。2014 年汉语方言语法调查框架与莱比锡标注系统高级研修班授课讲义。

亢世勇（2004）《面向信息处理的现代汉语语法研究》，上海：上海辞书出版社。

陆丙甫、屈正林（2010）语义投射连续性假说：原理和引申———兼论定语标记的不同功能基础，《语言学论丛》（第四十二辑），北京：商务印书馆，112-128 页。

王瑞晶（2010）语义地图：理论简介与发展史述评，《语言学论丛》（第四十二辑），北京：商务印书馆，81-111 页。

吴福祥（2009）从“得”义动词到补语标记 — 东南亚语言的一种语法化区域，《中国语文》第 3 期。

吴福祥（2011）多功能语素与语义图模型，《语言研究》第 1 期，页。

张　敏（2010）“语义地图模型”：原理、操作及在汉语多功能语法形式研究中的运用，《语言学论丛》（第四十二辑），北京：商务印书馆，3-60 页。

Anderson, Lloyd B. (1982). The “Perfect” as a Universal and as a Language-particular Category. In Paul J. Hopper (ed.), Tense-Aspect: Between Semantics & Pragmatics, pp.227-264. Amsterdam: Benjamins.

Boye , Kasper(2007). Semantic Maps and the Identification of Cross-linguistic Generic Categories: Evidentiality and its Relation to Epistemic Modality”. http://www.eva.mpg.de/lingua/conference/07-SemanticMaps/files/manuscripts.html.

Croft, William(2001). Radical Construction Grammar. Oxford: Oxford University Press.

Croft, William & Keith T. Poole(2008). Inferring Universals from Grammatical Variation: Multidimensional Scaling For Typological Analysis. Theoretical Linguistics 34.1, pp.1-37.

Cysouw, Michael(2007). Building Semantic Maps: the Case of Person Marking. In: Matti Miestamo & Bernhard Wälchli (eds)., New Challenges in typology: Broadening the horizons and redefining the foundations. Berlin: Mouton, P225-248.

de Haan, Ferdinand(2004). On Representing Semantic Maps. Ms. University of Arizona.

de Haan, Ferdinand(2007). Building a Semantic Map: Top-down Versus Bottom-up Approaches.http://www.eva.mpg.de/lingua/conference/07-SemanticMaps/files/manuscripts.html.

Haspelmath, Martin(1997). Indefinite Pronouns. Oxford: Clarendon.

Haspelmath, Martin(2003).The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In:M. Tomasello (ed.), The New Psychology of Language, vol. 2, New York: Erlbaum, 211-243.