Mini-Challenge 1 Report

林子玥 21110980025
施润叶 21110980026
黄占波 21210980036

1 可视化系统总览

1.1 新闻之间的关系及偏见

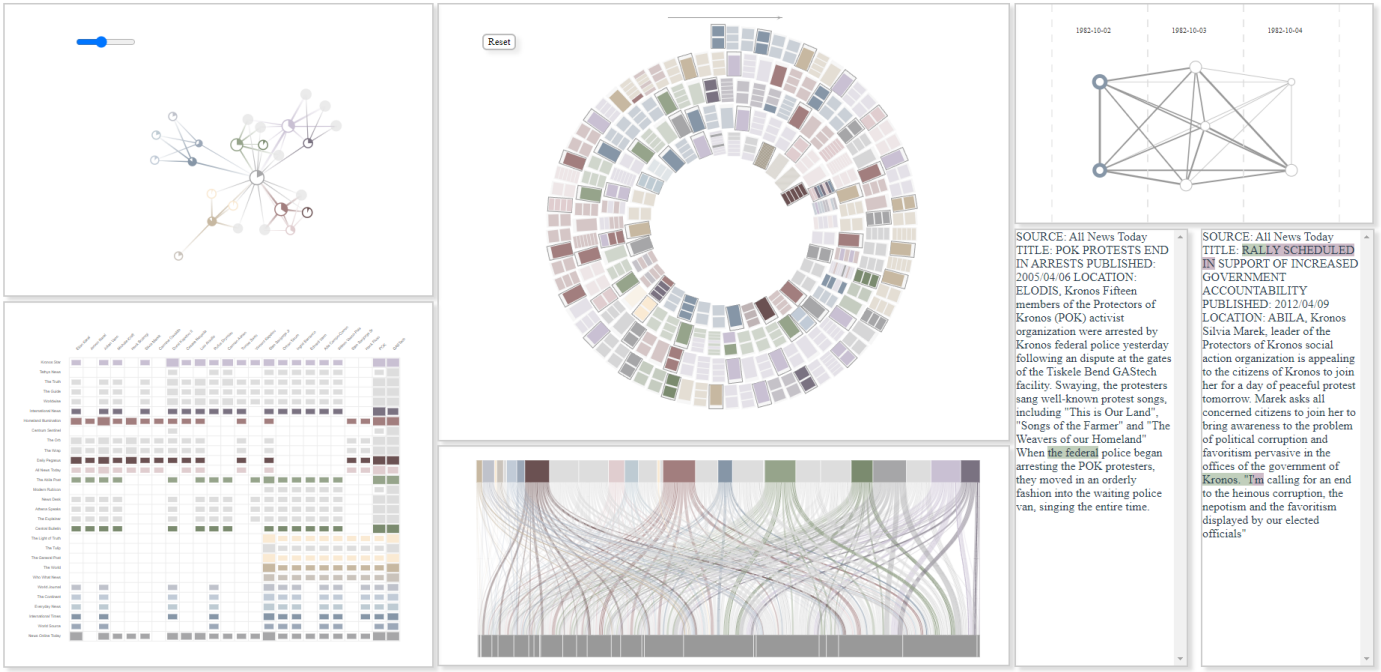


图 1 新闻之间的关系视图 1

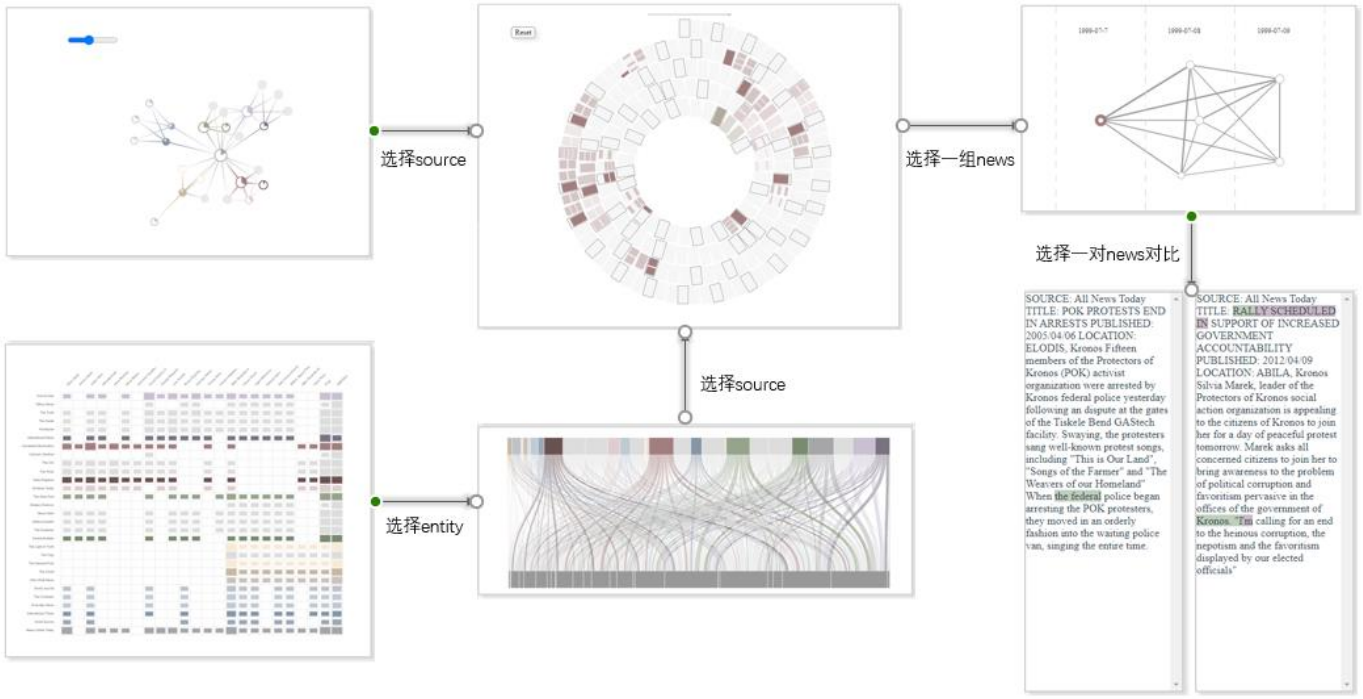


图 2 视图 1 交互流程

1.2 组织及成员之间的关系

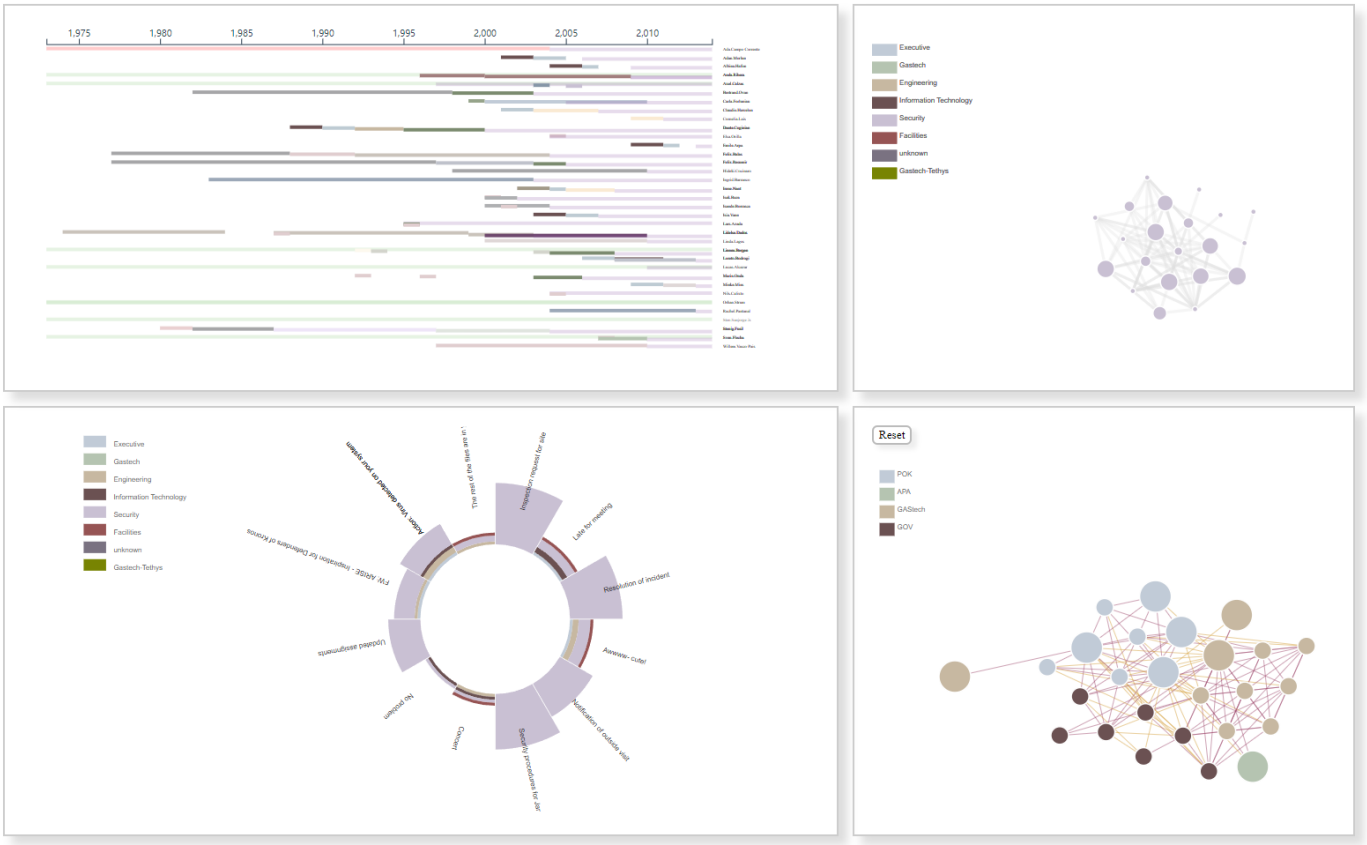


图 3 组织关系视图 2

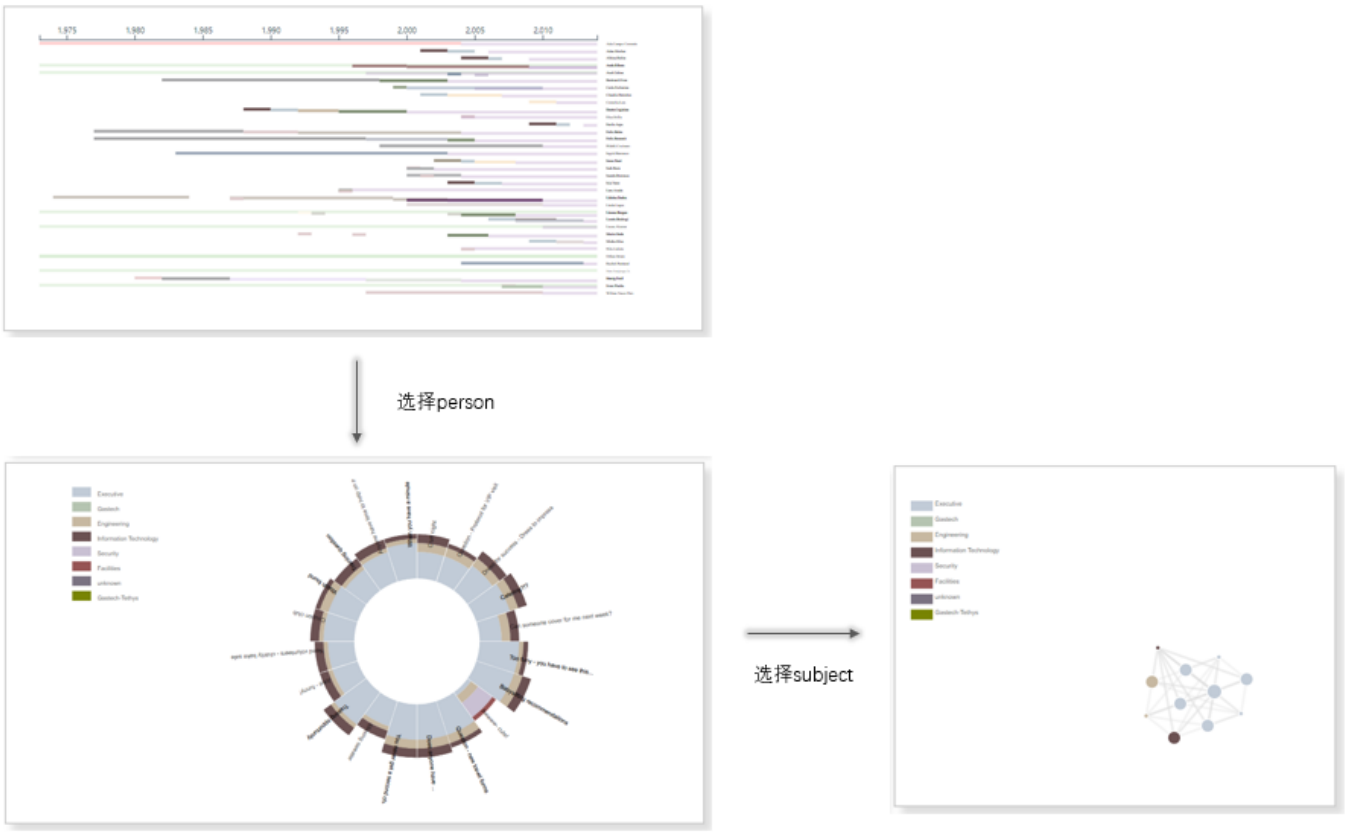


图 4 视图 2 交互流程

2 第一题

2.1 哪些新闻是原创的？哪些新闻是衍生的？

首先，我们利用文本分析 doctovec 的办法对新闻文本数据进行了处理，根据文本内容的重复程度计算了新闻文稿之间的相似性，并以新闻发布前后作为根据，选择最先发表的文稿作为原创新闻。根据新闻文稿之间的借鉴关系，我们再计算各个新闻社之间总的一个衍生关系。我么可以通过以下两个视图对新闻社及新闻文稿之间的关系有一个大致的了解。

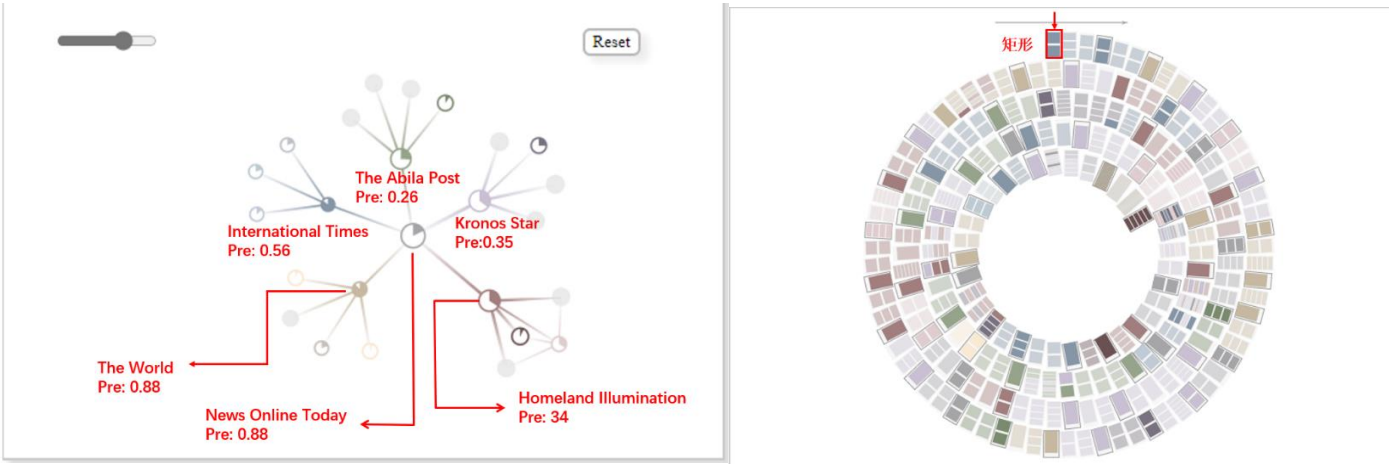


图 5 （左）原创新闻以及衍生新闻之间的关系：每个新闻机构用一个圆饼图表示，切片的大小表示为该新闻机构中原创内容的多少，即标示 Pre。（右）新闻通稿的时间顺序：每一个矩形代表一天，按照时间排序；矩形中每一个色块代表一篇新闻通稿，同一内容的新闻颜色相同，但原创新闻的透明度低；黑色框线表示与前一天相隔超过 7 天，前后新闻没有联系。

从图 5（左）中我们可以看出做原创新闻的主要有 International Times, The Abila Post, Homeland Illumination, The Kronos Start, The World 五家，而其他新闻都或多或少借鉴了这五家新闻社的内容，具体原创与衍生的关系如表 1 所示。其中 News Online Today 则是集中了五家的新闻。

表 1 原创与衍生新闻的关系

Kronos Star	**Homeland Illumination**	**The Abila Post**	**The World**	**International Times**
Tethys Nows	Centrum Sentinel	Modern Rubicon	The Tulip	World Journal
The Truth	The Orb	News Desk	The Light of Truth	The Continent
The Guicle	The Wrap	Athena Speaks	The General Post	Everyday News
Worldwise	Daily Pegasus	The Explainer	Who What News	News Online Today
International News	All News Today	Central Bulletin		World Source

为了了解具体新闻稿之间的借鉴关系，我们可以先通过点击对应新闻社的节点，会显示相应新闻社以及其邻居节点发布的新闻稿。通过点击具有衍生关系的一组新闻稿，得到的新的视图之间描述了新闻稿之间的借鉴关系以及相似程度。以 2013 年 9 月 2 日 The World 发布的一篇原创新闻稿为例，我们可以 The Tulip 和 The General Postfenbie 在 9 月 3 日和 9 月 4 日借鉴了这篇文章，并在当天各发布了两篇衍生新闻稿。而 The Tulip 在第二天发布的 X 与它的相似度最高，距离为 0.14。

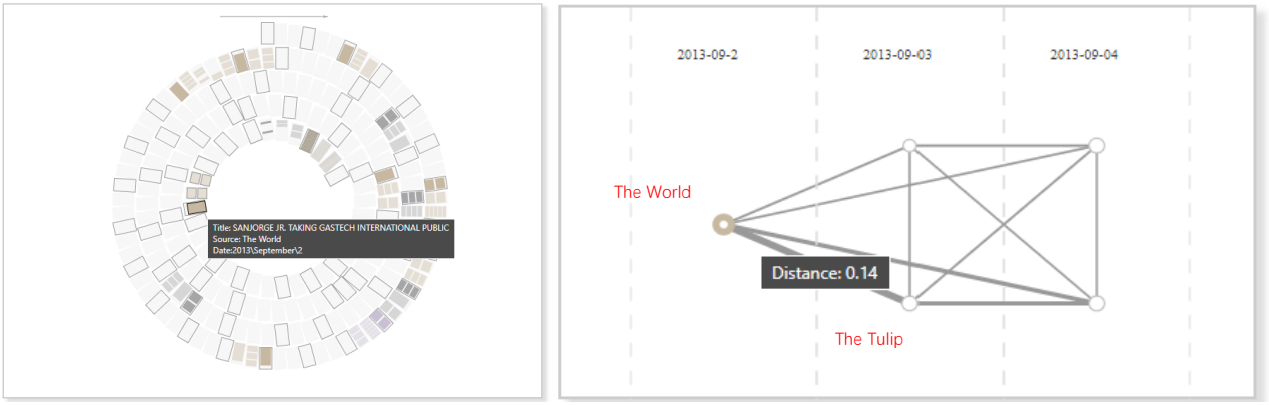


图 6 原创新闻与衍生新闻之间的相似性举例。右图每一个节点表示一个新闻文稿，节点与节点间的连线权重表示链接节点间的相似性

2.2 原创新闻与衍生新闻之间的关系

原创新闻与衍生新闻之间的详细关系可以通过比较新闻文稿来回答。我们认为一共可能有这样几种关系。

1) 抄袭

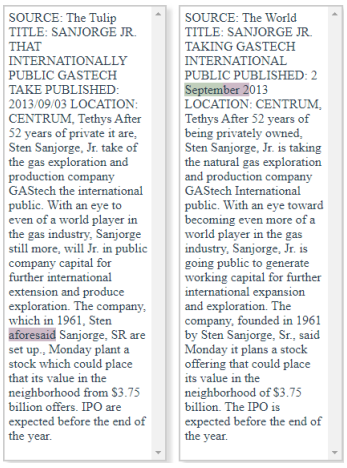


图 7 抄袭衍生，左边是衍生新闻，右边是原创新闻

其中红色的高亮表示这是对比另一篇新闻稿多余的部分，而绿色的高亮表示这是对比另一篇新闻稿修改的部分。我们可以发现这两篇内容基本一致。

2) 额外信息

图 8 额外信息衍生，左边是原创新闻，右边是衍生新闻

衍生新闻对比原创新闻增加了一些细节，比如医院床位数量这些信息的增加，但具体内容没有改变。

3) 后续报道

图 9 后续报道衍生，左边是原创新闻，右边是衍生新闻

原创新闻只报道到车祸的结果，而衍生新闻增加了调查的内容。

4) 情感修饰与观点叠加

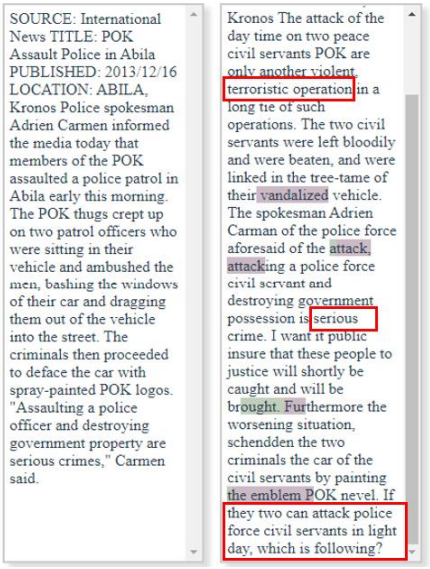


图 10 修饰衍生，左边是原创新闻，右边是衍生新闻

衍生新闻相对原创新闻来说描写得更加生动，在客观报道内容上加入了记者自己的观点，但具体内容换汤不换药。

3 第二题

3.1 在新闻中存在哪些偏见，举例

我们所涵盖的偏见包括两种类型的偏见。一种是包含实体的偏见，即新闻在关注的群体上可能有偏好；另一种是涵盖话题的偏见，即新闻在报道某一实体时更关注于哪方面的内容。

首先我们通过 Spacy 包中名词识别和命名实体识别，从新闻文稿中识别并根据实体在新闻稿中被提及的次数选出了来自 POK、GASTech 和 Government 的共 24 个实体，包含有 22 个人名，并对每家新闻社对不同人物的报道做了一个矩阵。

图 11 新闻机构对人物报道的偏见。其中原创新闻和衍生新闻用相似的色块表示，色块的大小表示对应实体被该新闻机构报道的次数

我们发现衍生于同一个原创新闻社的衍生新闻与原创新闻通常对人物的报道是非常相似的。比如以 The World 为主要来源的这组新闻机构只关注 GASTech 的成员，International Times 为主要来源的这组新闻机构也相比于 POK 和 Goverment 更加关注 GASTech；在这其中，被提及最多的 GASTech 成员是他们的 CEO Sten Sanorge Jr.。以 Homeland Illumination 为主要来源的这组新闻机构更关注 POK，而不太关注 GASTech 的成员；在 POK 中被他们所提及最多的是 Julian Vann，一个因为 GASTech 环境污染逝去的年轻女孩。而以 Kronos Star 和 The Abila Post 为主要来源的新闻机构对于 POK 成员的关注就相对于他们对于 Government 和 GASTech 成员的关注少一点，但总体对于三个组织的关注比较平均。

在点击每个实体之后，他右边这个视图会展示与这个实体（人名）相关的新闻机构和话题。机构与话题之间的报道关系我们用 Sankey Diagram 表现。

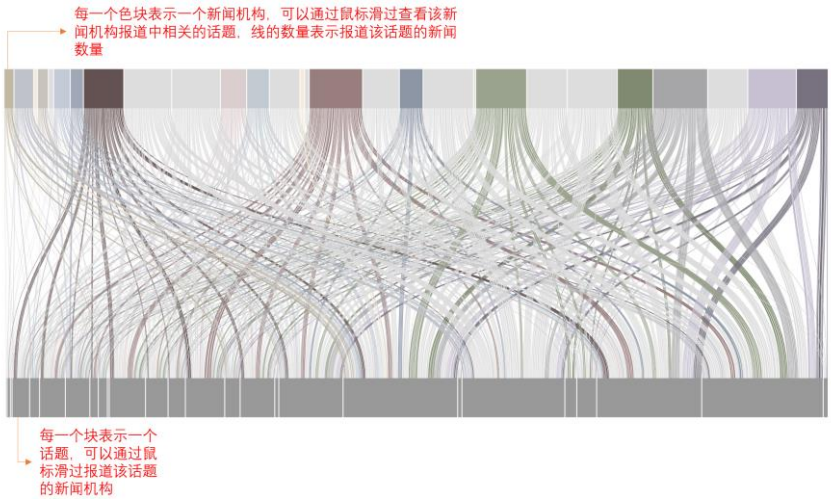


图 12 Sankey Diagram

我们发现每组新闻机构在报道同一个实体时也会带有话题的偏见。

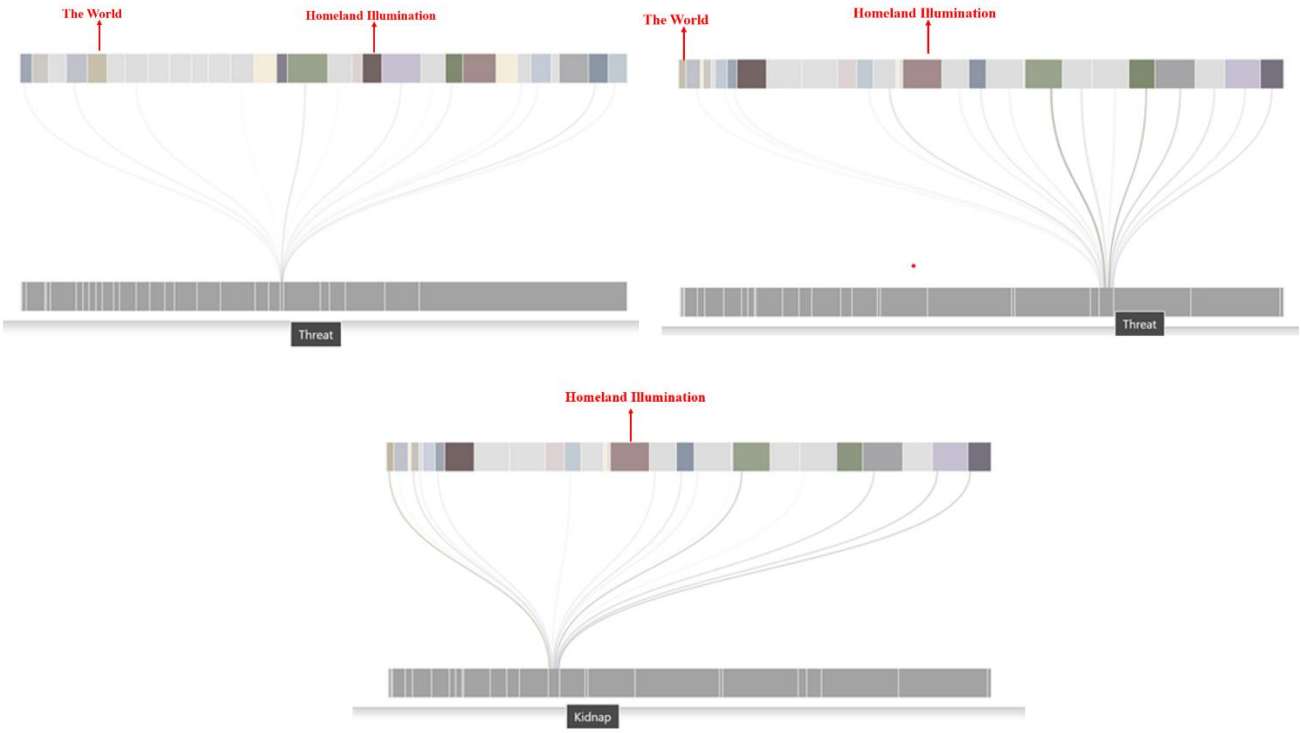
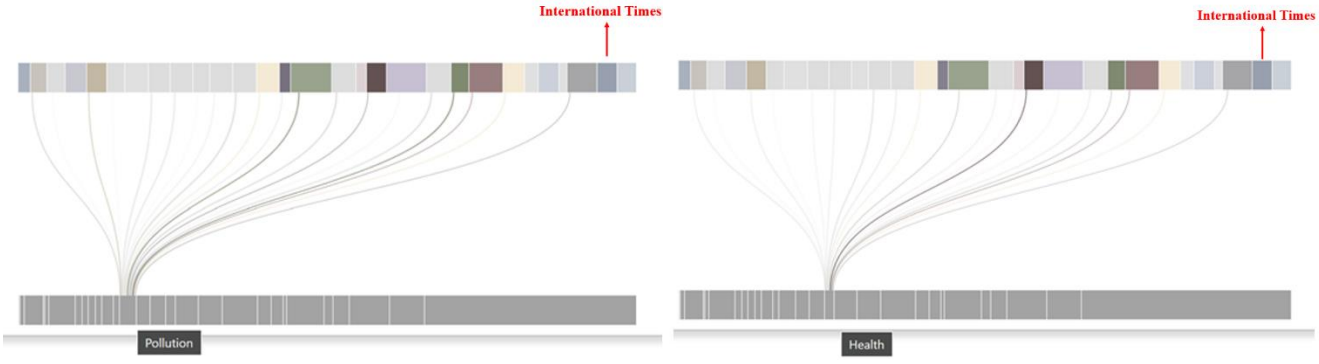


图 13 （上左）报道 GASTech 和 Threat 的机构。（上右）报道 POK 和 Threat 的机构。（下）报道 POK 和 Kidnap 的机构。

比如在报道 GASTech 的时候，只有 Homeland Illumination 和 The World 这两组新闻机构完全没有提及 Threat 的话题。而在提到 POK 的时候，Homeland Illumination 也没有提到 Kidnap 的话题，也就是对于 Homeland Illumination 来说，他们是不承认 POK 的成员绑架了 GasTech 的员工的说法，也没有这个倾向的。



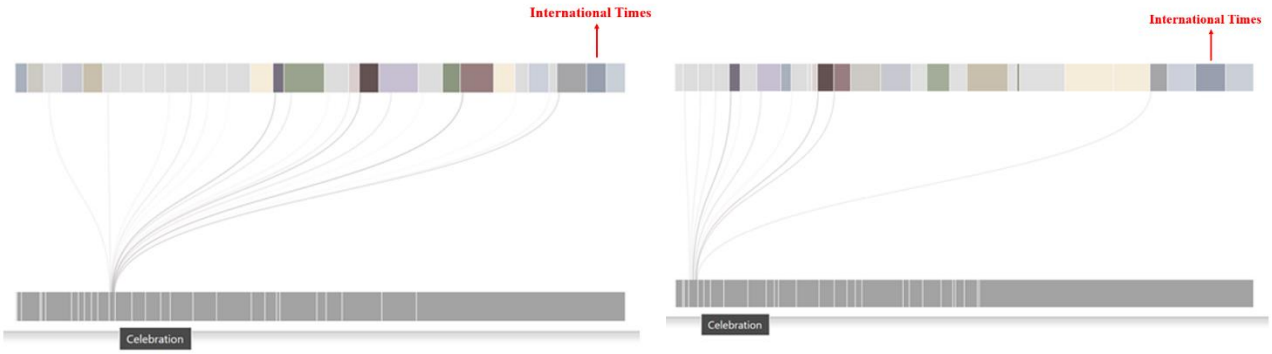


图 14 （左上）报道 GASTech 和 Pollution 的机构。（右上）报道 GASTech 和 Health 的机构。（左下）报道 GASTech 和 celebration 的机构。（右下）报道 GASTech 的 CEO Sten Jr.和 celebration 的机构。

而 International Times 这组新闻机构在提及 GASTech 的时候，没有涉及到 Pollution 和 Health 的话题，也没有涉及到 celebration 的话题。他们对待 GASTech 的态度就很模糊不清，但我们认为他们还是偏袒 GASTech 的，或者只是比较关注于绑架这起案件。

4 第三题

我们一共构建了四个视图，包括有个人简历信息、某个人的邮件往来信息、某一封具体邮件中的人员往来信息和从新闻中识别出来的人物关系信息。

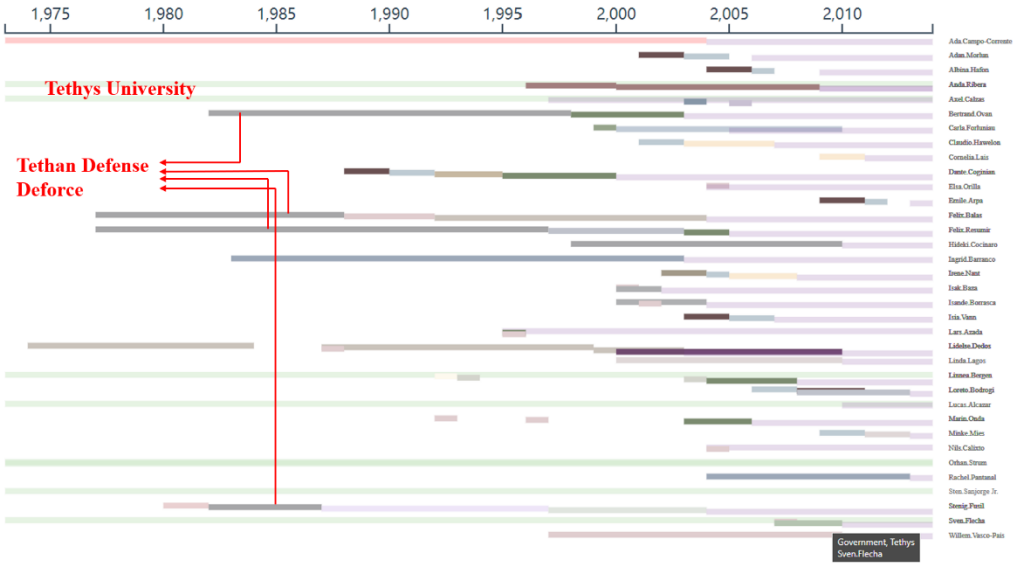


图 15 个人简历信息

每个色条代表一个人不同的一段经历，相同经历用一个颜色表示。我们发现 Sven Flecha 曾经在 2005-2010 年间在政府就职，他可能与政府有关系。另外，Anda Ribera 和 Axel Calzas 都曾经在 Tethys University 有很长的一段学习/任职经历，时间也比较重合，他们两可能很早就认识，关系比较好。而 Betrand Ovan 和 Felix Balas 和 Felix Resumir 和 Stenig Fusil 都曾经在 Tethan Defence Force 服役，且服役时间段有重合，他们也可能认识，特别是两个 Felix 的服役时间相同。

点击图 15 的人名后，我们可以看到他与别人的邮件往来

这里以 Isia Vann 为例。

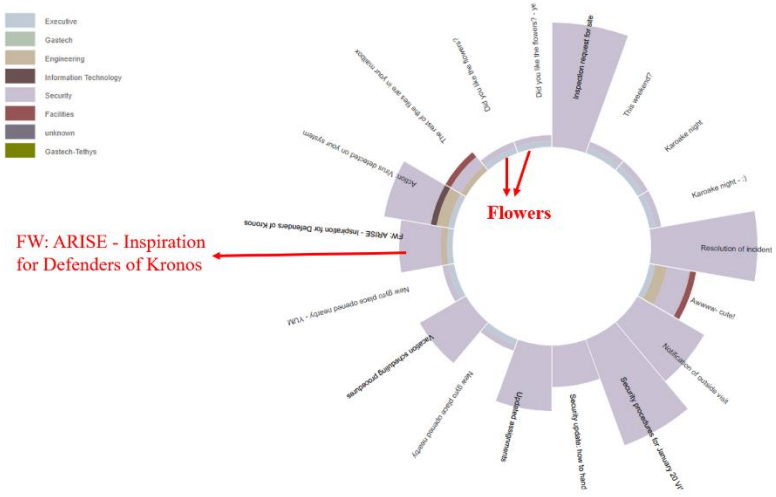


图 16 Isia 与他人的邮件往来

我们发现在 Isia Vann 的邮件里发现他跟别人经常讨论 flower，这可能是他涉毒的证据；并且信息部门提醒他在他的电脑里发现了病毒，而病毒主要在 GASTech 安保部门成员的电脑里发现。另外，Isia 的邮件中还有一封关于”FW: ARISE - Inspiration for Defenders of Kronos”的邮件，说明他是 POK 的成员。而与这份邮件有关的成员如图 17 所示。所以我们有理由怀疑 Hennie Osvaldo, Inga Ferro, Isia Vann, Loreto Bodrogi, Minke Mies, Rachel Pantanal, Ruscella Mies Haber 都是 POK 潜伏在 GASTech 中的成员。而 Isia 和 Loreto 都曾经在 Armed Forces of Kronos 服役。

图 17 （左）FW: ARISE - Inspiration for Defenders of Kronos 邮件的相关人员。（右）Armer Forces of Kronos 的服役记录

另外，我们还通过对新闻文稿进行情感识别，根据新闻文稿在不同组织成员中建立关系 Opponent 或者 Ally。通过初步观察，我们发现：GASTech 成员和 POK 成员间基本是敌对关系。

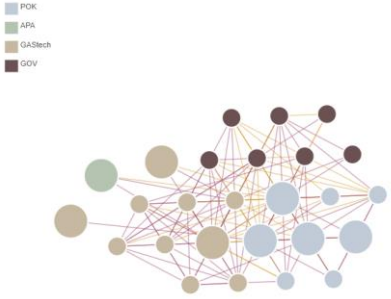


图 18 不同组织成员由新闻文稿导出的关系

除了 Isia Vann 和因 GASTech 环境污染逝去的 Julia Vann 是兄妹，Hank Fluss 和 POK 的创始人 Henk Bodrogi 是战友关系。而 Government 中的新旧关系也很复杂：旧主席 Luis Araullo 和旧卫生部部长 Cesare Nespola 支持 POK，认为他们能带动经济；而新的主席 Dorel Kapelou 和新的卫生部部长 Vincent Kapelou（叔侄关系）支持 GASTech，认为需要保护环境。他们形成了政府中的两股势力。

图 19 （左）Isia Vann 和 Julia Vann 的兄妹关系。（右）Hank Fluss 和 Henk Bodrogi 的战友关系

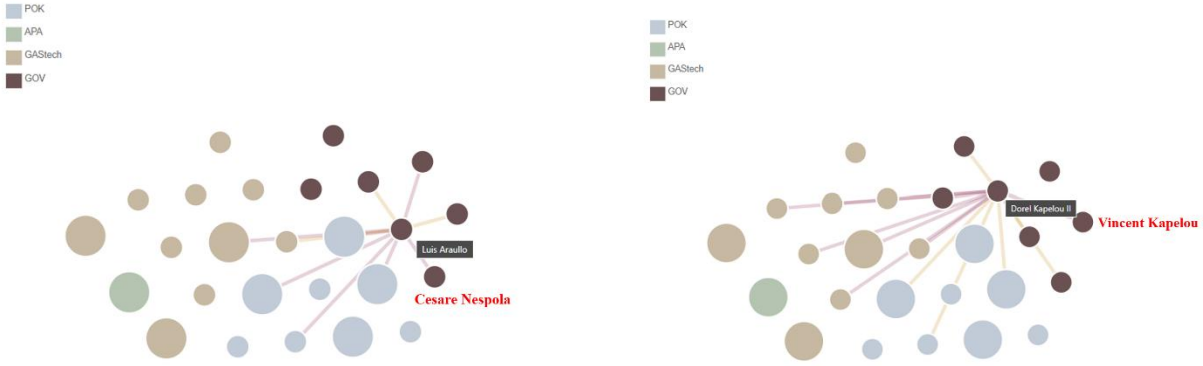


图 20（左）旧主席 Luis Araullo 的关系。（右）新主席 Kapelou 的关系

5　小组成员分工

表 2 小组成员分工

小组成员	分工
林子玥	可视化系统
施润叶	数据处理、分析、报告
黄占波	数据处理、报告