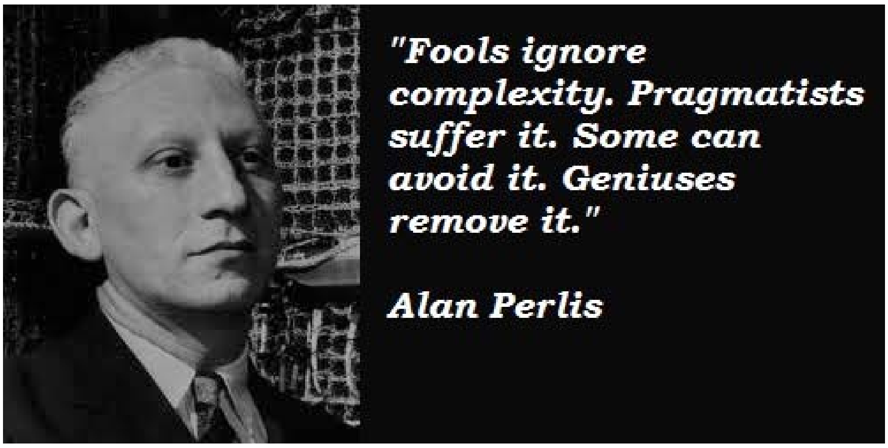
NP 完全性证明举例

陈卫东
chenwd@scnu.edu.cn
华南师范大学计算机学院
2021-11



https://www.brainyquote.com/authors/alan_perlis

一、多项式时间归约“$\le_p$ ”的作用

1. 算法设计视角

如果有 **B**$\le_p$ **A**，则问题 **B** 可转化为问题 **A** 来求解。

特别地，若 **A** 有多项式时间算法，则 **B** 也有。即，若 **A**∈**P**，则 **B**∈**P**。

2. 计算复杂性视角

如果有 **A**$\le_p$ **B**，则问题 **B** 的难度不低于问题 **A** 的难度。

特别地，若 **A** 是 **NP-hard** 问题，则 **B** 也是 **NP-hard** 问题。

二、NP 完全性的证明技巧

1. **B** 是 **NP-complete** 问题的证明思路：

（1）B∈NP；

（2）A $\le_p$ B，其中 A 是某个 NP-complete 问题。（这一条成立表明 B 是 NP-hard 问题）

*[注] 通常找结构相似的问题进行归约。例如，图的 k-color (k>3)问题可用图的 3-color 着色问题归约，4-D 匹配问题可用 3-D 匹配问题归约等。

2. 如果问题 **A** 是问题 **B** 的特例（实际上有 **A** $\le_p$ **B**），且 **A** 是 **NP-complete** 问题，则 **B** 是 **NP-complete** 问题。

例如，划分问题是子集和问题的特例，子集和问题是 0-1 背包问题的特例。

三、若干例子

注：例子来自教材《**Algorithm Design**》中的第 **8** 章练习题，解答来自网络资料的整理。

1. 你正在为一个小的高技术公司维护高度保密的计算机系统，该系统正在用于一件很敏感的工作. 为了保证这个系统不被用于任何违法的目的，公司安装了一个记录软件用来记录它的所有使用者访问的 IP 地址. 假设每一位使用者在任意给定的一分钟内至多访问一个 IP 地址. 该软件写一个记录文件，对每一位使用者 u 和每一分钟 m 记录一个值 $I(u,m)$. 这个值等于使用者 u 在第 m 分钟内访问的 IP 地址（如果有的话）；如果在第 m 分钟使用者 u 没有访问任何 IP 地址，则置 $I(u,m)$ 为空符号 $\perp$.

公司管理部门刚才获悉昨天有人用系统对一个很远的站点发动了复杂的攻击. 攻击的实现是通过在 t 个连续的分钟内访问 t 个不同的 IP 地址：攻击在第 1 分钟访问地址 i_1，在第 2 分钟访问地址 i_2，…，直到在第 t 分钟访问地址 i_t.

谁应该对此负责？公司检查了记录，奇怪地发现没有一位使用者 u 在相应的时间内访问有关的地址，换句话说，没有 u 使得对从 1 到 t 的每一分钟 m，$I(u,m)=i_m$.

于是，问题变成：是否有一个 k 位使用者的小集团共同实现了攻击？设使用者的子集 S，如果对从 1 到 t 的每一分钟 m，至少有一位使用者 $u\in S$ 使得 $I(u,m)=i_m$，则我们称 S 是一个*可疑集团*（换句话说，每一个 IP 地址在相应的时间内至少被这个集团中的一位使用者访问过）.

可疑集团问题问：给定所有的值 $I(u,m)$ 和数 k，存在不超过 k 位使用者的可疑集团吗？

解 首先，显然**可疑集团**属于 $\mathcal{NP}$. 如果给我们出示了使用者集合 S，我们可以核实 S 至多有 k 位使用者，并且从 1 到 t 的每一分钟 m，S 中至少有一位使用者访问过 IP 地址 i_m.

现在我们要找一个已知的 NP 完全问题并把它归约到**可疑集团**. 虽然有许多特征（使用者，分钟，IP 地址），但是很显然它是一个覆盖问题（按照本章描述的分类）：必须解释所有 t 个可疑的访问，并且只允许有限位（k 位）使用者做这些事. 一旦确定这是一个覆盖问题，自然地打算把**顶点覆盖**或**集合覆盖**归约到它. 为此，把它的大多数复杂的特征忽略掉是有益的，只留下骨架式的特征用来描述**顶点覆盖**或**集合覆盖**.

我们把**顶点覆盖**归约到这个问题. 在**顶点覆盖**中，必须覆盖每一条边并且只允许 k 个顶点. 在**可疑集团**中，必须“覆盖”所有的访问并且只允许 k 位使用者. 这种平行的关系强烈地提示我们，给定一个由图 $G=(V,E)$ 和数 k 构成的**顶点覆盖**的实例，要构造一个**可疑集团**的实例，应该在这个实例中用使用者表示 G 的顶点，用可疑的访问表示边.

假设**顶点覆盖**实例中的图 G 有 m 条边 $e_1,\cdots,e_m$，其中 $e_j=(v_j,w_j)$. 如下构造**可疑集团**的实例. 对 G 的每一个顶点构造一位使用者，对每一条边 $e_t=(v_t,w_t)$ 构造一个分钟 t（因此总共有 m 分钟）. 在第 t 分钟，与 e_t 的两个端点相关的两位使用者访问 IP 地址 i_t，其余所有的使用者不访问任何 IP 地址. 最后，攻击由在第 $1,2,\cdots,m$ 分钟分别访问地址 $i_1,i_2,\cdots,i_m$ 构成.

下述断言表明**顶点覆盖** $\leqslant_p$ **可疑集团**，从而完成**可疑集团**是 NP 完全的证明. 实例的构造与原有的**顶点覆盖**实例的关系如此密切，以至于下面的证明完全是直截了当的.

断言 8.28 *在构造的实例中，存在大小不超过 k 的可疑集团当且仅当图 G 有大小不超过 k 的顶点覆盖.*

证 首先，假设 G 有大小不超过 k 的顶点覆盖 C. 考虑**可疑集团**实例中对应的使用者集合 S. 对从 1 到 m 的每一个 t，C 中至少有一个元素是边 e_t 的端点，从而 S 中对应的使用者访问 IP 地址 i_t. 因此，集合 S 是一个可疑集团.

反过来，假设存在大小不超过 k 的可疑集团 S，考虑 G 中对应的顶点集合 C. 对从 1 到 m 的每一个 t，S 中至少有一位使用者在第 t 分钟访问 IP 地址 i_t，从而 C 中对应的顶点是边 e_t 的一个端点. 因此，集合 C 是一个顶点覆盖. ■

2. 某商店为了分析它的顾客的行为，要经常维护一个 2 维数组 A，其中行对应它的顾客，列对应它出售的商品. 元素 $A[i,j]$是顾客 i 购买商品 j 的数量.

下面是一个很小的这样的数组 A

	液体清洁剂	啤酒	尿布	猫用干草
Raj	0	6	0	3
Alanis	2	3	0	0
Chelsea	0	0	0	7

商店可能想用这些数据做下面的事情. 如果顾客子集 S 中的任意两位顾客从来没有买过相同的商品(即，对每一种商品，S 中至多有一位顾客买过)，则称 S 是**兴趣互异的**. 兴趣互异的顾客集合可能是有用的，例如，它可能集中了市场调查的研究对象.

兴趣互异的子集问题定义如下：给定如上定义的一个 $m \times n$ 数组 A 和一个数 $k \leqslant m$，有至少有 k 位顾客的兴趣互异的子集吗?

证明**兴趣互异的子集**是 NP 完全的.

The problem is in $\mathcal{NP}$ because we can exhibit a set of k customers, and in polynomial time is can be checked that no two bought any product in common.

We now show that *Independent Set* $\leq_P$ *Diverse Subset*. Given a graph G and a number k, we construct a customer for each node of G, and a product for each edge of G. We then build an array that says customer v bought product e if edge e is incident to node v. Finally, we ask whether this array has a diverse subset of size k.

We claim that this holds if and only if G has an independent set of size k. If there is a diverse subset of size k, then the corresponding set of nodes has the property that no two are incident to the same edge — so it is an independent set of size k. Conversely, if there is an independent set of size k, then the corresponding set of customers has the property that no two bought the same product, so it is diverse.

9. 考虑下述问题. 你正在管理一个通信网络，用有向图 $G=(V,E)$ 作为这个网络的模型. 有 c 个用户有兴趣使用这个网络. 对每一个 $i=1,2,\cdots,c$，用户 i 发出一条请求，要求预订 G 中的一条路径 P_i 用于传输数据.

你们有兴趣接受尽可能多的这种路径请求，但要服从下述限制：如果接受 P_i 和 P_j，那么 P_i 和 P_j 不能共享任何结点.

于是，**路径选择**问题问：给定有向图 $G=(V,E)$，请求 $P_1, P_2, \cdots, P_c$(每一个 P_i 必须是 G 中的一条路径)以及数 k，问至少选择 k 条路径并且任意两条选出的路径都不共享结点是可能的吗?

Path Selection is in NP, since we can be shown a set of k paths from among $P_1, \ldots, P_c$ and check in polynomial time that no two of them share any nodes.

Now, we claim that *3-Dimensional Matching* $\leq_P$ *Path Selection*. For consider an instance of *3-Dimensional Matching* with sets X, Y, and Z, each of size n, and ordered triples $T_1, \ldots, T_m$ from $X \times Y \times Z$. We construct a directed graph $G = (V, E)$ on the node set $X \cup Y \cup Z$. For each triple $T_i = (x_i, y_j, z_k)$, we add edges (x_i, y_j) and (y_j, z_k) to G. Finally, for each $i = 1, 2, \ldots, m$, we define a path P_i that passes through the nodes $\{x_i, y_j, z_k\}$, where again $T_i = (x_i, y_j, z_k)$. Note that by our definition of the edges, each P_i is a valid path in G. Also, the reduction takes polynomial time.

Now we claim that there are n paths among $P_1, \ldots, P_m$ sharing no nodes if and only if there exist n disjoint triples among $T_1, \ldots, T_m$. For if there do exist n paths sharing no nodes, then the corresponding triples must each contain a different element from X, a different element from Y, and a different element from Z — they form a perfect three-dimensional matching. Conversely, if there exist n disjoint triples, then the corresponding paths will have no nodes in common.

Since *Path Selection* is in NP, and we can reduce an NP-complete problem to it, it must be NP-complete.

(Other direct reductions are from Set Packing and from Independent Set.)

Reduce Independent Set to Path Selection as follows. Let (G,k) be an instance of the Independent Set problem. Let $G=(V,E)$ with $|V| = n$ and $|E| = m$.

The reduction creates a new directed graph $G^*= (V^*,E^*)$ and n paths $P_1,P_2,...,P_n$ such that G has an independent set of size k if and only if there are k paths in $P_1,P_2,...,P_n$ that are node-disjoint in G^*. The graph G^* has m vertices, one corresponding to each edge of G. We let a_e denote a vertex in G^* where e is an edge in E. We make G^* a complete directed graph which means that there is a directed edge between every pair of vertices $(a_e, a_{e'})$. It remains to define the paths. The paths correspond to vertices in G. For each vertex $i \in V$ (of G), there is a path P_i. Let $e_{j1},e_{j2},...,e_{jh}$ be the edges incident to i in G (in some arbitrary order). Then the path P_i is $e_{j1} \rightarrow e_{j2} \rightarrow ... \rightarrow e_{jh}$; note that this is a valid path since G^* is a complete directed graph. It can be seen that G^* and the paths $P_1,P_2,...,P_n$ can be constructed in polynomial time from G. One can show that $S=\{i_1,i_2,...,i_t\}$ is an independent set in G iff the paths $P_{i1},P_{i2},...,P_{it}$ are node-disjoint. The reason is that if $e=(i,j)$ is an edge in E then P_i and P_j both contain the vertex a_e in G^*. And conversely if P_i and P_j contain a node a_e in G^* then $e=(i,j)$ in G.

10. 你在 WebExodus 的几位朋友最近在给几家公司做几项咨询工作，这些公司在维护大众可登录的大型网站（由于契约问题他们不能说出给哪家公司做咨询）. 他们遇到下述**广告策略**问题.

一个公司带着一个网站的图来找他们，我们用有向图 $G=(V,E)$ 表示这个网站的图. 这个公司还提供了网站用户的一组 t 个典型的搜索轨迹，我们用 G 中的有向路径 $P_1, P_2, \cdots, P_t$ 表示这些轨迹（即，每一条 P_i 是 G 中的一条路径）.

这个公司要求 WebExodus 回答下述问题：给定 G，路径集合 $\{P_i\}$ 和数 k，最多在 G 的 k 个结点放置广告，使得每一条路径 P_i 上至少有一个结点放置广告，这是可能的吗？我们称这个问题为**广告策略**问题，它以 $G, \{P_i \mid i=1,\cdots,t\}$ 和 k 为输入.

你的朋友估计这个问题的好算法能够使他们都富起来. 不幸的是，事情绝不是这么简单.

(a) 证明**广告策略**是 NP 完全的.

(a) We'll say a set of advertisements is "valid" if it covers all paths in $\{P_i\}$. First, *Strategic Advertising* (SA) is in NP: Given a set of k nodes, we can check in $O(kn)$ time (or better) whether at least one of them lies on a path P_i, and so we can check whether it is a valid set of advertisements in time $O(knt)$.

We now show that *Vertex Cover* $\leq_P$ SA. Given an undirected graph $G = (V, E)$ and a number k, produce a directed graph $G' = (V, E')$ by arbitrarily directing each edge of G. Define a path P_i for each edge in E'. This construction involves one pass over the edges, and so takes polynomial time to compute. We now claim that G' has a valid set of at most k advertisements if and only if G has a vertex cover of size at most k. For suppose G' does have such a valid set U; since it meets at least one end of each edge, it is a vertex cover for G. Conversely, suppose G has a vertex cover T of size at most k; then, this set T meets each path in $\{P_i\}$ and so it is a valid set of advertisements.

18. 要求你帮助几位组织理论家分析团体决策的数据. 特别地，他们正在研究一张由一个特殊的政府政策委员会所做决定的数据表，并且试图确定是否能够识别出一个由该委员会中有影响的成员组成的小集合.

该委员会是这么工作的. 它有 n 个成员 $M=\{m_1,\cdots,m_n\}$，并且在过去的一年里投票表决了 t 项不同的议题. 对每一个议题，每一位成员可以投票"赞成"、"反对"或"弃权". 最后的结果是，如果赞成票数严格地多于反对票数(弃权票不计在任何一方)，则委员会对这项议题呈递肯定的决议，否则提交否定的决议.

现在我们有一张很大的表，它由委员会的每一位成员对每一项议题投的票组成. 考虑下述定义.

我们说成员的子集 $M'\subseteq M$ 是**决定性的**，如果只考虑 M' 中成员投的票，委员会对每一项议题的决定也会是一样的.（换句话说，M' 中成员的投票结果与整个委员会的投票结果在所有的议题上都是一样的.）可以把这样一个子集看做影响整个委员会行为的"内部集团".

这里有一个问题：给定每一位成员对每一个议题投的票和参数 k，我们希望知道是否存在不超过 k 位成员组成的决定性子集. 我们说这是**决定性的子集**问题的一个实例.

例 有 4 位委员会成员和 3 个议题. 我们正在寻找大小不超过 $k=2$ 的决定性子集，投票情况如下：

议题 #	m_1	m_2	m_3	m_4
议题 1	赞成	赞成	弃权	反对
议题 2	弃权	反对	反对	弃权
议题 3	赞成	弃权	赞成	赞成

因为成员 m_1 和 m_2 构成一个决定性的子集，所以这个实例的答案是"yes".

证明**决定性的子集**是 NP 完全的.

The problem *Decisive Subset (DS)* is in NP because we can check in polynomial time that a given subset of committee members is of size at most k, and that its voting outcome is the same as that of the whole committee.

Now we show that *Vertex Cover* $\leq_P$ *DS*. Given a graph $G=(V,E)$ and bound k, we create an issue I_e for each edge e, and a committee member m_v for each node v. If $e=(u,v)$, then we have members u and v vote "yes" on issue I_e, and all other committee members abstain. Note that the voting outcome by the whole committee on all issues is "yes." We now ask whether there is a decisive subset of size at most k.

If there is a decisive subset S of size at most k, then it must lead to an affirmative decision for each issue. In particular, this means that for each edge e, S must include at least one of the members m_u or m_v, and so the corresponding set of nodes will constitute a vertex cover in G of size at most k.

If there is a vertex cover C of size at most k, then for each edge $e=(u,v)$, at least one of u or v will be in C. For the set S of members corresponding to C, the voting outcome will thus be "yes" on each issue I_e, so S is decisive.

34. 20 世纪 70 年代包括 Mark Granovetter 和 Thomas Schelling 在内的数学社会科学领域中的研究人员开始尝试开发人类某些类型的集体行为的模型. 为什么某些时尚流行，而另一些过时了？为什么某些新的技术发明被广泛地采用，而另一些仍旧只有少数的使用者？用什么动力学解释有时（尽管只是很稀少地）在愤怒的人群中出现骚乱和抢劫行为？他们提出这些都是瀑布过程的例子，在瀑布过程中个别人的行为受到他或她的朋友们的行为的极大地影响，因而如果少数人发起这种过程，它可能传给越来越多的人，最终产生非常广泛的影响. 我们可以认为这个过程像疾病或者流言传播一样，从一个人传给另一个人.

模型最基本的样式如下. 有一个基本的行为（比如，打冰球、拥有一部手机、参加骚乱），并且在任何时间每一个人是这个行为的采用者、或者不采用者. 我们用有向图 $G=(V,E)$ 表示全体居民，图中的结点对应人. 如果 v 对 w 的行为有影响，即 v 采用这个行为有助于促使 w 也采用这个行为，则有一条边 (v,w). 每一个人 w 有一个给定的阈值 $\theta(w)\in[0,1]$，阈值 $\theta(w)$ 的涵义是：在任何时间当有边到 w 的所有结点中至少有 $\theta(w)$ 是这个行为的采用者时，结点 w 也变成这个行为的采用者.

注意到具有低阈值的结点更容易被说服采用这个行为，而具有高阈值的结点更保守些. 阈值 $\theta(w)=0$ 的结点 w 立即采用这个行为，而不受朋友们的影响. 最后，对无进入边的结点，我们约定：如果阈值等于 0 则他们成为采用者，否则他们不可能成为采用者.

给定这个模型的一个实例，我们可以如下模拟行为的传播.

```
开始，设所有 θ(w)=0 的结点 w 为采用者
   （其他所有的结点开始时为不采用者）
Until  采用者集合没有变化：
    For  每一个不采用者 w 同时做：
        If  有边到 w 的所有结点中至少有 θ(w) 是采用者 then
            w 成为采用者
        Endif
    Endfor
End
输出最终的采用者集合
```

注意到这个过程一定终止，因为一共只有 n 个人，每一次循环至少有一个人成为新的采用者.

最近几年，市场和数据挖掘的研究人员已经看到可以如何用这种模型来研究新产品成功中的“口碑”效应（所谓病毒市场现象）. 想法是这样的，涉及到的行为是使用一种新产品，我们可能能够说服居民中少数关键人物试用这种产品，并且希望引发尽可能大的流行.

具体地说，假设我们选择结点集合 $S\subseteq V$ 并重新令 S 中的结点的阈值为 0（通过说服他们试用产品，确保他们是采用者）. 然后运行上面描述的过程，并且看最终的采用者集合有多大. 把这个最终的采用者集合的大小记作 $f(S)$（注意，因为它自然地依赖于 S 的选择，所以我们把它写成 S 的函数）. 可以把 $f(S)$ 看做集合 S 的影响，因为它反映了当在 S 上“播种”时行为传播得有多广.

如果我们正在销售一种产品，那么目标是寻找一个小的集合 S，使其影响 $f(S)$ 尽可能的大. 于是，定义**影响最大化**问题如下：给定有向图 $G=(V,E)$，每一个结点有一个阈值，以及参数 k 和 b，问有不超过 k 个结点的集合 S 使得 $f(S)\geqslant b$ 吗？

证明**影响最大化**是 NP 完全的.

例 假设图 $G=(V,E)$ 有 5 个结点 $\{a,b,c,d,e\}$，4 条边 (a,b)，(b,c)，(e,d)，(d,c)，并且所有结点的阈值都等于 2/3. 那么，由 G 以及 $k=2$ 和 $b=5$ 定义的**影响最大化**实例的答案是 yes. 可以选择 $S=\{a,e\}$，这将导致另外 3 个结点也变成采用者(在这里，这是使 S 符合要求的唯一选择. 例如，如果选择 $S=\{a,d\}$，则 b 和 c 会变成采用者，而 e 不能变成采用者. 如果选择 $S=\{a,b\}$，则 c，d 和 e 没有一个能变成采用者).

First we need to make sure that the problem is in NP. This is as in the previous cases easy to see. Given an initial set of size k, we can just run the influence algorithm mentioned in the assignment and obtain a final influence set, and then we can check whether the size of the influence set is at least b. The algorithm is polynomial, since at each iteration we are either terminating or at least adding another vertex to the already influenced list.

Now we need to prove that this problem is also NP-complete. We reduce from vertex cover. Let $G=(V,E)$ be a graph and k be a parameter for the instance of the vertex cover problem. We will create a node for each of the vertices v in G as well as for each of the edges e in G (we will call them vertex-nodes and edge-nodes). There is an edge (v,e) if v is one of the ends of the edge e. All the thresholds on nodes are $1/2$. We want to claim that a vertex cover of size k corresponds to a initial set of size k with the resulting influence set of size $(k+m)$ where m is the number of edges in G, and vice versa.

One direction is easy - if there is a vertex cover of size k then this vertex cover will infect all the edges, therefore we will have a influenced set of size $(k+m)$.

Now if there is an influenced set of size $(k+m)$ from an original set of k nodes, we claim that there is a vertex cover of size k. Notice that if there is an edge-node in the k size original set, then we can replace it by one of the vertex nodes adjacent to it, and the edge will get infected immediately. So therefore doing this operation repeatedly, we will get a set of size k consisting entirely of vertices that infects $(k+n)$ nodes. But this set would be a vertex cover, since each edge would have to have a neighbor among the original vertex nodes, otherwise it would not be infected.