**实验数据：**

1. **特点：**
   1. binary （01）型的关联数据，主要包括三部分，以 DDI预测为例的话有已知的DDI相关矩阵 D （药物跟药物的关联关系）、DTI相关矩阵 R （药物跟蛋白质的关联关系）和 PPI 关联矩阵 T （蛋白质跟蛋白质的关联关系）。本质上是 network 型的数据，但是需要的是3个相关的网络。
2. **预计使用的实验数据有：**
   1. DT （药物-蛋白质）相关数据
   2. GP （基因-表型）相关数据
   3. UI （用户-商品）相关数据
   4. UB （用户-事务）相关数据

具体如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类型 | 组成 | 数据量 | 格式 | 源文件 |
| DT | D, R, T | 1000\*1000左右 | 数据点对 | DDI.zip （备用的：drug\_target） |
| GP | G, R, P | 1500\*1500左右 | 文本型 | Mus\_gp |
| UI | U, R, I | 13000\*13000左右（截取部分使用） | Matlab数据文件 | Douban.mat |
| UB | U, R, B | 16000\*16000左右（截取部分使用） | Matlab数据文件 | Yelp.mat |

注意：

1. Mus\_gp 的数据是小鼠的基因跟表型关联数据，原始数据是从MGI 和 OMIM数据集中下载的文本文件，需要进行一些预处理和对齐，特别是MPO，即表型-表型关联数据，因为关联关系是树形的，需要截取4-5层的关联，另外这些数据一个月或者几个月会更新一次，有时候需要进行实验数据上的更新。
2. DDI.zip是 **Systematic Prediction of Pharmacodynamic Drug-Drug Interactions through Protein-Protein-Interaction Network** 这篇文章作者收集处理的数据，目前相关链接可能已经失效了，不过之前下载的原数据还在。
3. Drug\_target 是根据DTI预测使用的数据集处理后得到的，因此它的PPI数据并不是真实的 PPI关联，而是根据相似度得到的伪关联
4. Douban.mat和Yelp.mat这两份数据是根据shi chuan 的论文 **Semantic Path based Personalized Recommendation on Weighted Heterogeneous Information Networks** 所获得的，数据虽然已经处理成mat格式，但是本质上只是将原始数据转化成了mat中的cell格式，具体应用的时候还需要进行截取（原始关联数太多了，计算量过于庞大）和处理。
5. 虽然network型的开源数据很多，但是符合实验条件的数据还是不易获得的，后期需要寻找新的数据集，可以考虑 shi chuan, sunyizhou等人的工作，基于metapath的文章大多是使用异质信息网络，因此比较容易找到契合的数据。

**实验设置：**

1. **实验数据说明：**
   1. 对使用的数据进行说明，可以在每个实验小结分别说明，也可以总结到一个表格中
   2. 说明包括，数据的组成（数据由哪几部分组成），数据的大小（数据维度），关联数，数据来源，数据介绍（可以跟数据的组成一起说明）
2. **交叉检验：**
   1. 使用10-folds或者 5-folds的交叉检验的方式来进行实验，具体可以根据实验效果来进行选择（因为10-folds和5-folds都是比较常用的）
   2. 进行交叉检验时，要将已知的关联矩阵的对应行列同时置为0，作为训练数据，置为0的那些行列作为测试集 （因为行列是对称的）
3. **评价方式：**
   1. 基于link prediction 的实验中，预测的**准确率**跟**召回率**是最重要的评价指标，根据实验的要求，需要使用 AUCn, AUPRn （可以添加precisio\_n和 recall\_n），n根据一定的区间进行划分选取 （如10， 20， 50， 100， 200， all）。
   2. 相应的图表有 AUCn、AUPRn，precision\_n, recall\_n等的对应表格和对应柱状图，结果数据应该是10次交叉检验结果的平均值和对应方差
   3. 可以人工添加部分噪声（以一定比率添加），再进行实验，相应的表格和图就是实验结果基于噪声比例的变化，指标如上（此时n可以取固定值）。
4. **参数分析：**
   1. 主函数部分有2块，一个是单边流形，另一个是双边流形，需要对双边流形的参数进行讨论，双边流形的参数中讨论的主要是先验约束alpha和beta，用对应的折线图来展示
   2. 正则项部分如果使用了低秩约束或者其它特殊约束，也要对其参数进行讨论，对应的符号是gamma
   3. 如果先验约束中使用了indicator矩阵W的话，还要对W的值进行说明，因为W对应的是不同先验知识的权重，主要说明的就是mu 和rho（具体参考论文）
5. **验证分析（Validation）:**
   1. 作为评价方式的补充，可以增加现实数据的validation部分，早期的DDI预测中，可以利用online tools （如 DDI checker， DrugBank等）进行验证，PPI预测也可以使用kegg进行验证（因为 PPI的数据是不断更新的，可以用旧的关联作为来预测新的关联）
   2. Social network的数据目前没有比较好的validation方式，后期可以考虑用别的方式进行补充分析
6. **对比方法：**
   1. 对比方法的选取有比较大的选择空间，已有的对比方法中有一些可以延续使用，如：标签传播算法 （Label Propagation Prediction of Drug-Drug Interactions Based on Clinical Side Effects，注意的是它在进行传播矩阵归一化是使用了BBS方法，具体参照论文），SVM（或者是LR等方法，代表的是classification类方法 ），随机游走算法RWR，distance learning类方法（如 皮尔森相关系数， cosine相似度等， 相关方法较多，需要综合考虑）
   2. 增加的新方法可以从 Maximum likelihood methods 和 Probabilistic models两方面入手，具体参考论文 **Link prediction in complex networks: A survey**