Working doc

1 Graphlet kernel

Graphlets. The graphlet sampling kernel decomposes graphs into graphlets (i.e. small subgraphs with k nodes) and counts matching graphlets in the input graphs. Let $H_1, H_2, \ldots, H_{N_k}$ be the set of size-k graphlets. Here we have two choices: either we distinguish graphlets that are isomorphic as different, in which case $N_k = 2^{\frac{k(k-1)}{2}}$, or we don't, in which case N_k is still exponential in k but still lower. The first choice has the disadvantage of involving a much higher N_k, but the advantage of not requiring to solve the graph isomorphism problem to identify graphlets. Clever procedures exist for small k, but in general no polynomial algorithm is known.

We denote by $\phi_k^{hist.}$ the function on size-k graphs that identifies the corresponding graphlet and serves to construct the corresponding histogram:

$$\phi_k^{hist.}(F) = \left[1_{(F=H_i)}\right]_{i=1}^{N_k} \in \{0,1\}^{N_k}$$

In other words, $\phi_k^{hist.}$ puts a 1 in the coordinate i if $F = H_i$, and 0 otherwise. Here, equality between F and H_i is to be taken up-to-isomorphism or not, depending on the choice made above.

Sampling and graphlet kernel. For a graph G, let $S_k(G)$ be a *sampling procedure* that randomly extract a subgraph of size k from G. The simplest choice is independent node sampling, but many other methods exist. For a graph G and a sampling procedure S_k, let $f_G \in [0,1]^{N_k}$ be a vector such that its i^{th} entry is equal to the frequency of occurrence of graphlet i in G when sampling with S_k:

$$f_G = [\mathbb{P}(S_k(G) = H_i)]_{i=1}^{N_k} = \mathbb{E}_{F \sim S_k(G)}[\phi_k^{hist.}(F)]$$

For two graphs G, G', the graphlet kernel is then defined as:

$$k(G, G') = f_G^\top f_{G'} \tag{1}$$

which naturally involves an associated Euclidean metric $d_k(G, G') = \|f_G - f_{G'}\|_2$.

Subsampling. Exhaustive enumeration of graphlet is very expensive. Since there are $\binom{k}{n}$ size-k subgraphs in a graph (when using independent node sampling), computing the feature vector for a graph of size n requires $O(n^k)$ time.

So, in practice, given a graph G, we sample independently n subgraphs $F_1, \ldots, F_n$ using the sampling process S_k. Then, f_G is simply approximated with an empirical average

$$f_G \approx \frac{1}{n} \sum_{j=1}^{n} \phi_k^{hist.}(F_j) = \hat{\mathbb{E}}[\phi_k^{hist.}(F)] \tag{2}$$

Using simple concentration inequalities, it can be showed [**?**] that by sampling a fixed number of graphlets the empirical distribution of graphlets will be sufficiently close to their actual distribution if the graph.

Issues. We identify two issues with the traditional graphlet kernel:

- For general k, the function $\phi_k^{hist.}$ itself is expensive to compute. Unless smart procedures can be employed for small k, there is no other way than go through the entire graphlet list (and solve or not the graph isomorphism on top of that), and N_k is at least exponential in k.
- the inner product (1) and its associated metric do not take into account a notion of *similarity* between the graphlet themselves, they just compare the frequency counts for each graphlet, independently from the other.

We adress both question by replacing the function ϕ_k with an efficient, randomized high-dimensional mapping.

2 Random features, MMD

Random Features Consider a p.d. kernel k between objects $x \in \mathcal{X}$. Random features for kernel is based on the following expression:

$$k(x, y) = \mathbb{E}_{\omega \sim \Lambda} \xi_\omega(x)^* \xi_\omega(y) \tag{3}$$

for some family of mapping $\xi_\omega : \mathcal{X} \to \mathbb{C}$ indexed by a parameter $\omega \in \mathbb{R}^d$ (often called "frequency" because of Random *Fourier* features) distributed according to some distribution Λ.

For instance, based on Bochner Theorem: a continuous translation-invariant kernel $k(x, y) = k(x - y)$ on $\mathbb{R}^d$ is positive definite if and only if $k(\delta)$ is the Fourier transform of a non-negative measure, i.e. it satisfies (3) with $\xi_\omega(x) = e^{i\omega^\top x}$. It is not hard to see that complex exponential can also be replaced with cosines with dithering: $\xi_{\omega,b}(x) = \sqrt{2}\cos(\omega^\top x + b)$ where $b \sim \mathcal{U}([0, 1])$.

The RF methodology simply consists in approximating (3) with an empirical average: given $\omega_1, \ldots, \omega_m$ drawn *iid* from Λ, we have

$$k(x, y) \approx \frac{1}{m} \sum_{\ell=1}^{m} \xi_{\omega_\ell}(x)^* \xi_{\omega_\ell}(y) = \phi(x)^* \phi(y) \tag{4}$$

where

$$\phi(x) = \frac{1}{\sqrt{m}} \left[\xi_{\omega_\ell}(x)\right]_{\ell=1}^{m}$$

contains the random mappings of x. Hence we have replaced k with a *linear* kernel between the $\phi(x)$.

Preprocessing For graph(let)s, it may be useful to include to preprocessing function $\gamma(F) \in \mathbb{R}^d$, potentially invariant by permutation, before computing the traditional random features:

$$\xi_\omega(G) = \tilde{\xi}_\omega(\gamma(F))$$

where $\tilde{\xi}_\omega$ are traditional RF on $\mathbb{R}^d$, eg, random fourier features.

Mean kernel and MMD The mean kernel methodology allows to *lift* a kernel from a domain $\mathcal{X}$ to a kernel on *probability distributions* on $\mathcal{X}$. Given a base kernel k and two probability distribution P, Q, it is defined as

$$k(P, Q) = \mathbb{E}_{x \sim P, y \sim Q} k(x, y) \tag{5}$$

In words, the mean kernel is just the expectation of the base kernel with respect to each term. The associated Euclidean metric is called the *Maximum Mean Discrepancy* (for quite obscure reasons), and is naturally defined as

$$MMD(P, Q) = \sqrt{k(P, P) + k(Q, Q) - 2k(P, Q)} \tag{6}$$

Warning: note that $k(P,P) = \mathbb{E}_{x\sim P, x'\sim P} k(x,x') \neq \mathbb{E}_{x\sim P} k(x,x)$.

If the kernel has the form (3), then it is immediate that we have

$$MMD(P,Q)^2 = \mathbb{E}_\omega \Big(|\mathbb{E}_P \xi_\omega(x) - \mathbb{E}_Q \xi_\omega(x)|^2 \Big) \tag{7}$$

Given data $x_1, \ldots, x_n$ drawn *iid* from P and $y_1, \ldots, y_n$ drawn *iid* from Q, the kernel (5) can naturally be approximated by

$$k(P,Q) \approx \frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i, y_j) \tag{8}$$

and the corresponding approximate MMD is

$$MMD(P,Q) \approx \sqrt{\frac{1}{n^2} \sum_{i,j=1}^{n} k(x_i,x_j) + k(x_i,x_j) - k(x_i,y_j) - k(x_j,y_i)}$$

Note that many variants are possible.

MMD and random features Mean kernel goes especially well with random features. Combining (4) and (8), it is not hard to see that using RF the mean kernel can be further approximated by

$$k(P,Q) \approx \frac{1}{n^2} \sum_{i,j=1}^{n} \phi(x_i)^* \phi(y_j) = \left(\frac{1}{n} \sum_i \phi(x_i) \right)^* \left(\frac{1}{n} \sum_i \phi(y_i) \right) \tag{9}$$

So the computation can be drastically improved by first computing the *averaged random features* (also called random *generalized moments*, also called *sketch*) $\frac{1}{n}\sum_i \phi(x_i)$, and taking a linear kernel between them. The corresponding MMD is then just the Euclidean metric between the averaged random features

$$MMD(P,Q) \approx \| \frac{1}{n} \sum_i \phi(x_i) - \frac{1}{n} \sum_i \phi(y_i) \|_2$$

MMD for discrete distribution Note that, for a discrete space of objects $H_1, \ldots, H_N$ with discrete probability distributions $P = [P_1, \ldots, P_N]$ and Q on them, the mean kernel (5) takes a particular form

$$k(P,Q) = \sum_{i,j=1}^{N} P_i Q_j k(H_i, H_j)$$

3 Combining the two

One can see the link with graphlet sampling, where f_G is the (discrete) probability distribution of the graphlets. If we define $k(F,F') \approx \phi(F)^* \phi(F')$ where ϕ is a random feature map that replaces ϕ_k^{hist}, then the feature map (2) is exactly what appears in (9). So, now, all the game becomes to find a good feature map $\phi(F)$ for graphlets. The induced MMD metric between graph is the MMD between graphlets probability distributions f_G:

$$d(G,G') = MMD(f_G, f_{G'}) = \sqrt{k(f_G,f_G) + k(f_{G'},f_{G'}) - 2k(f_G,f_{G'})} \approx \| \frac{1}{n} \sum_i \phi(F_i) - \frac{1}{n} \sum_i \phi(F'_i) \|_2$$

where F_i are graphlets drawn from G and F'_i are graphlets drawn from G'.

It has been shown that the approximation error in (2) in L_1 norm is small with high probability (already in Hashem report).

For us, it is interesting to see how much, given two graphs G and G', $\|\frac{1}{n}\sum_i \phi(F_i) - \frac{1}{n}\sum_i \phi(F_i')\|_2$ is close to $d(G, G')$.

Lemma 1 (todo)**.** *Let G and G' be two graphs. Draw F_i (resp. F_i') iid with $S_k(G)$ (resp. $S_k(G)$) and define $y = \frac{1}{n}\sum_i \phi(F_i)$ (resp. $y' = \frac{1}{n}\sum_i \phi(F_i')$) . We have*

$$\mathbb{P}(|\|y - y'\|^2 - d(G, G')^2| \geq \varepsilon) \leq ...$$

Proof. We decompose the proof in two steps.

Step 1: infinite n, finite m. Using Hoeffding's inequality, prove that:
$d(G, G')^2$ is close to $\frac{1}{m}\sum_{j=1} |\mathbb{E}_{F\sim f_G}\xi_{\omega_j}(F) - \mathbb{E}_{F'\sim f_{G'}}\xi_{\omega_j}(F')|^2 = \|\mathbb{E}_{F\sim f_G}\phi(F) - \mathbb{E}_{F'\sim f_{G'}}\phi(F')\|^2$. Use equation (7).
Taking into assumption that when we approximate the graphlet kernel using random features, then the Λ distribution in (3) satisfies that for each simple graph x we have:

$$0 \leq \xi_\omega(x) \leq 1, \forall \omega \sim \Lambda \tag{10}$$

This is a reasonable assumption since the lifting function ϕ_k^{hist} of a k-graphlet kernel includes the normalized frequency of occurrences of each graphlet of size k in the graph x. Thus, in this case and based on this assumption we can make use of Hoeffding's inequality that states:

Lemma 2 (Hoeffding's inequality)**.** *Let $(x_1, \ldots, x_m)$ be independent random variables such that the variable x_i is strictly bounded by the interval $[a_i, b_i]$, and let $\overline{X} = \frac{1}{m}\Sigma_{i=1}^m x_i$ then we have:*

$$Pr(|\mathbb{E}\overline{X} - \overline{X}| \geq \epsilon) \leq 2\ exp(-\frac{2m^2\epsilon^2}{\Sigma_{i=1}^m (b_i - a_i)^2}) \tag{11}$$

In our case, and for a finite number of random features (m) we have the variables $x_j = |\mathbb{E}_{F\sim f_G}\xi_{\omega_j}(F) - \mathbb{E}_{F'\sim f_{G'}}\xi_{\omega_j}(F')|^2$ are independent (we assume here that when the variables $w_{j\in 1,\ldots,m}$ are independent then for every graph F the variables $\xi_{\omega_{j\in\{1,\ldots,m\}}}(F)$ are independent too) and bounded by the interval $[0, 1]$ too, thus by straight forward application one can prove:

$$Pr(\Big|\frac{1}{m}\sum_{j=1}^m |\mathbb{E}_{F\sim f_G}\xi_{\omega_j}(F) - \mathbb{E}_{F'\sim f_{G'}}\xi_{\omega_j}(F')|^2 - \mathbb{E}_\omega|\mathbb{E}_P\xi_\omega(x) - \mathbb{E}_Q\xi_\omega(x)|^2\Big| \geq \epsilon) \leq 2\ e^{-2m\epsilon^2} \tag{12}$$

Step 2: finite n and m. Show that for any *fixed* set of random features ω_j, we have $\|\mathbb{E}_{F\sim f_G}\phi(F) - \mathbb{E}_{F'\sim f_{G'}}\phi(F')\|$ close to $\|\frac{1}{n}\sum_i \phi(F_i) - \frac{1}{n}\sum_i \phi(F_i')\|$. For this, a version of Hoeffding's inequality for *vectors* might be useful, see for instance Lemma 4 in Appendix A of "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning" by Rahimi and Recht.
Let us consider a fixed set of random variables $\{\omega_j\}_{j\in\{1,\ldots,m\}}$ drawn independently from Λ, thus the random features map of a graph F equals: $\phi(F) = \frac{1}{\sqrt{m}}\left[\xi_{\omega_j}(F)\right]_{j=1}^m$.
For every graph G , let $F_1, \ldots, F_n$ be n random subgraphs drawn independently from G, we clearly have:

$$\mathbb{E}_{F\sim f_G}\phi(F) = \mathbb{E}(\ \frac{1}{n}\sum_i \phi(F_i)\ ) \tag{13}$$

Here we will assume that we deal with non-sparse and Large-scale Graph G, so that sub-graph F_i cannot be predicted knowing subgraphs $F_1, \ldots, F_{i-1}$. i.e. we can consider the sub-graphs $(F_1, \ldots, F_n)$ (and thus $\phi(F_i)$) independent random variables. What should be noticed now to be used later is that $\forall F \sim f_G, \phi(F)$is in a ball $\mathcal{H}$ of radius $M = \frac{\sqrt{m}}{\sqrt{m}} = 1$.

Lemma 3. *let $X = x_1, \ldots, x_n$ be iid random variables in a ball $\mathcal{H}$ of radius M centered around the origin in a Hilbert space. Denote their average by $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then for any $\delta > 0$, with probability at lest $1 - \delta$,*

$$\|\overline{X} - \mathbb{E}\overline{X}\| \leq \frac{M}{\sqrt{n}}(1 + \sqrt{2\ log\frac{1}{\delta}} \tag{14}$$

Proof. Defining the function $f(x) = \|\overline{X} - \mathbb{E}\overline{X}\|$, and $\widetilde{X} = x_1, \ldots, \widetilde{x}_i, \ldots, x_n$ to be a copy of X with the ith element replaced by an arbitrary element of $\mathcal{H}$, we can prove using the triangle inequality:

$$|f(X) - f(\widetilde{X})| = \Big|\|\overline{X} - \mathbb{E}\overline{X}\| - \|\overline{\widetilde{X}} - \mathbb{E}\overline{X}\|\Big| \leq \|\overline{X} - \overline{\widetilde{X}}\| \leq \frac{\|x_i - \widetilde{x}_i\|}{n} \leq \frac{2M}{n} \tag{15}$$

Therefor, $f(X)$ is insensitive to the ith component of X, $\forall i \in \{1, \ldots, n\}$ which is an important requirement to apply McDiarmid's inequality on f.
To bound the expectation of f, we use the familiar identity about the variance of the average of *iid* random variables:

$$\mathbb{E}\|\overline{X} - \mathbb{E}\overline{X}\|^2 = \frac{1}{n}(\mathbb{E}\|x\|^2 - \|\mathbb{E}x\|^2) \tag{16}$$

Also:

$$\mathbb{E}f(X) \leq \sqrt{\mathbb{E}f^2(X)} = \sqrt{\mathbb{E}\|\overline{X} - \mathbb{E}\overline{X}\|^2} \leq \frac{M}{\sqrt{n}}$$

This bound for the expectation of f and McDiamid's inequality give:

$$Pr_x\Big[f(X) - \frac{M}{\sqrt{n}} \geq \epsilon\Big] \leq Pr_x\Big[f(X) - \mathbb{E}f(X) \geq \epsilon\Big] \leq exp\Big(-\frac{n\epsilon^2}{2M^2}\Big) \tag{17}$$

Which is equivalent to (14) by setting $\delta = exp(-\frac{n\epsilon^2}{2M^2})$ and solving for ϵ. □

Now back to Eq. (13) and its corresponding assumptions that we made, we can directly apply lemma 3 (and more especially Eq.(17)) to get that:

$$Pr(\|\mathbb{E}_{F\sim f_G}\phi(F) - \frac{1}{n}\sum_i \phi(F_i)\ \| \geq \frac{1}{\sqrt{n}} + \epsilon) \leq e^{-\frac{n\epsilon^2}{2}} \tag{18}$$

Now applying the triangle inequality again yields:

$$\Big|\|\mathbb{E}_{F\sim f_G}\phi(F) - \mathbb{E}_{F'\sim f_{G'}}\phi(F')\| - \|\frac{1}{n}\sum_i \phi(F_i) - \frac{1}{n}\sum_i \phi(F_i')\|\Big| \leq$$
$$\|(\mathbb{E}_{F\sim f_G}\phi(F) - \frac{1}{n}\sum_i \phi(F_i)) - (\mathbb{E}_{F'\sim f_{G'}}\phi(F') - \frac{1}{n}\sum_i \phi(F_i'))\| \leq$$
$$\|\mathbb{E}_{F\sim f_G}\phi(F) - \frac{1}{n}\sum_i \phi(F_i)\| + \|\mathbb{E}_{F'\sim f_{G'}}\phi(F') - \frac{1}{n}\sum_i \phi(F_i')\|$$

Thus, since the two variables $\|\mathbb{E}_{F\sim f_G}\phi(F) - \frac{1}{n}\sum_i \phi(F_i)\|$ and $\|\mathbb{E}_{F'\sim f_{G'}}\phi(F') - \frac{1}{n}\sum_i \phi(F_i')\|$ are independent (as a direct result of our aforementioned assumptions), $\forall \epsilon > 0$ we have:

$$Pr(\|\mathbb{E}_{F\sim f_G}\phi(F) - \frac{1}{n}\sum_i \phi(F_i)\| \geq \frac{1}{\sqrt{n}} + \frac{\epsilon}{2}, \|\mathbb{E}_{F'\sim f_{G'}}\phi(F') - \frac{1}{n}\sum_i \phi(F_i')\| \geq \frac{1}{\sqrt{n}} + \frac{\epsilon}{2}) =$$
$$Pr(\|\mathbb{E}_{F\sim f_G}\phi(F) - \frac{1}{n}\sum_i \phi(F_i)\| \geq \frac{1}{\sqrt{n}} + \frac{\epsilon}{2})\ Pr(\|\mathbb{E}_{F'\sim f_{G'}}\phi(F') - \frac{1}{n}\sum_i \phi(F_i')\| \geq \frac{1}{\sqrt{n}} + \frac{\epsilon}{2}) \leq e^{-\frac{n\epsilon^2}{4}}$$

finally, we get as a straight result from above:

$$Pr\Big(\Big|\|\mathbb{E}_{F\sim f_G}\phi(F) - \mathbb{E}_{F'\sim f_{G'}}\phi(F')\| - \|\frac{1}{n}\sum_i \phi(F_i) - \frac{1}{n}\sum_i \phi(F_i')\|\Big| \geq \frac{2}{\sqrt{n}} + \epsilon\Big) \leq e^{-\frac{n\epsilon^2}{4}}$$

and that is true for any fixed set of random variables $\{\omega_j\}_{j\in\{1,\dots,m\}}$ drawn independently from Λ.

Since it is valid for any fixed set of random features, it is also valid with *joint* probability on random features and samples, by the *law of total probability* (complex, we'll talk about it later).

Step 3 Using a *union bound*, conclude.

□

RIP. We may prove information preservation guarantees using the *Restricted Isometry Property* (RIP). The RIP is, somehow, a *uniform* version of (Step 1 of) Lemma 1. The features y we compute can be seen as a random embedding of f_G. If we have enough features, then all the information in f_G is contained in y, in the sense that there exists an *Instance Optimal Decoder* that can retrieve f_G from y, with some error (measured in the MMD). This is all the more true if f_G is well-approximated by a *sparse* vector, that has many coefficients close to 0.