A note on single-linkage equivalence

Andreas W.M. Dress a,∗, Taoyang Wu a, XiaoMing Xu b

a CAS-MPG Partner Institute and Key Lab for Computational Biology, 320 Yue Yang Road, 200031 Shanghai, China
b Department of Computer Science & Engineering, IJPL, Shanghai Key Lab, Fudan University, Shanghai, China

ARTICLE INFO

Article history:
Received 19 November 2009
Accepted 23 November 2009

Keywords:
Edge-weighted graphs
Single-linkage clustering
Maximal/minimal spanning trees
Single-linkage equivalence
Edge equivalence
Ultra-similarities
Ultra-metrics
Subdominant ultra-metrics

ABSTRACT

We introduce the concept of single-linkage equivalence of edge-weighted graphs, we apply it to characterise maximal spanning trees and “ultra-similarities”, and we discuss how it relates to the popular single-linkage clustering algorithm.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

In this note, we consider edge-weighted graphs defined on a fixed vertex set V, i.e., we consider triples G = (V, E, ω : E → R) given by specifying, in addition to the vertex set V, an edge set E = E ≤ (V ≤ 2) and a weight map ω = ω : E → R : [u, v] ← ω(u, v). In the context of the popular single-linkage clustering algorithm, such graphs are used to derive non-overlapping hierarchical cluster systems built up from “sufficiently similar” elements of V, interpreting the value ω(u, v), for any two distinct elements u, v ∈ V, as a measure of their (relative) similarity. Relying on Count Orlovski’s party principle “The friends of my friends are my friends”, one considers in single-linkage clustering, for any graph G as above and any threshold t in R ∪ {−∞}, the partition π(>)G(>t) of V into the various G-clusters of weight t, that is, the various connected components of the graph G(>t) with vertex set V and edge set E(>t) := {e ∈ E : ωC(e) > t}, the set of all those edges of G whose weight exceeds t.

It was observed 40 years ago (cf [1]) that single-linkage clustering can be performed very efficiently using maximal spanning trees. However, the closely related concept of single-linkage equivalence (of edge-weighted graphs) to be introduced in this note seems to have escaped attention in this context so far. However, as we will see below, it provides means to explore this well-known clustering procedure from a more abstract (or structural) point of view. More specifically, given a finite set V, we denote by W(V) the set of all edge-weighted graphs G with vertex set V, we define two graphs G, G′ ∈ W(V) to be single-linkage equivalent, if π(>)G(>t) = π(>)G′(>t) holds for every t ∈ R in which case we will also write G ≈ G′, and we denote, for any G ∈ W(V), by WW(V) the set of all graphs W(V) that are single-linkage equivalent to G. Clearly, W is an equivalence relation on W(V), in particular, WW(G) = WW(G′) holds for any graph G′ ∈ WW(G).

Next, we define a partial order “≰” on W(V) by defining G ≰ G′ for any two graphs G and G′ in W(V) if and only if EC is a subset of E_{C′} and ωC(e) ≤ ω_{C′}(e) holds for all e ∈ EC.
Clearly, denoting, for any two graphs \(G_1, G_2\) in \(W(V)\), by \(G_1 \wedge G_2\) the graph with edge set \(E := E_{G_1} \cap E_{G_2}\) and weight map \(\omega : e \in E : e \mapsto \min(\omega_{G_1}(e), \omega_{G_2}(e))\), we have \(G \not\sim G_1\) and \(G \not\sim G_2\) for any further graph \(G \in W(V)\) if and only if \(G \not\sim (G_1 \wedge G_2)\) holds.

We also note that, associating to every graph \(G \in W(V)\) the map

\[
\omega_G^* : \binom{V}{2} \rightarrow \mathbb{R} \cup \{-\infty\} : \{u, v\} \mapsto \begin{cases} 
\omega_G(u, v) & \text{if } \{u, v\} \in E, \\
-\infty & \text{otherwise,}
\end{cases}
\]

sets up a one-to-one correspondence between \(W(V)\) and the set \((\mathbb{R} \cup \{-\infty\})^{\binom{V}{2}}\) of all maps from \(\binom{V}{2}\) into \(\mathbb{R} \cup \{-\infty\}\) such that (i) \(\omega_G^* \wedge \omega_G = \min(\omega_G^*, \omega_G)\) holds for any two graphs \(G_1, G_2 \in W(V)\), and (ii) \(G \not\sim G'\) holds for any two graphs \(G, G' \in W(V)\) if and only if \(\omega_G^*(e) \leq \omega_{G'}^*(e)\) holds for all \(e \in \binom{V}{2}\)—a correspondence that will be used below again and again.

Finally, we define the map \(\omega_G^*\) associated to a graph \(G \in W(V)\) to be an ultrametric if the associated map \(\omega_G^*\) satisfies the ultra-similarity inequality, i.e., if

\[
\omega_G^*(u, v) \geq \min(\omega_G^*(u, w), \omega_G^*(w, v))
\]

holds for any three distinct vertices \(u, v, w\) in \(V\).

In this note, we will characterise, for any graph \(G \in W(V)\), the \(\not\sim\)-maximal and -minimal graphs in \(\forall(G)\) in terms of ultra-similarities and maximal spanning trees, first stating (in Section 2) then (in Section 4) establishing the pertinent results after having collected some fairly obvious, but rather useful simple facts in Section 3. And in the last section, we collect some simple applications and comments, including a brief discussion of the corresponding dual results regarding minimal spanning trees and subdominant ultra-metrics.

2. Five propositions

Continuing with the notations and definitions introduced above, we now consider a fixed edge-weighted graph \(G\) in \(W(V)\) to which we refer in our five main results below:

**Proposition 2.1.** There exists a unique \(\not\sim\)-maximal graph \(\overline{G}\) in the single-linkage equivalence class \(\forall(G)\) of \(G\).

**Proposition 2.2.** \(G\) coincides with the graph \(\overline{G}\) if and only if \(\omega_G^*\) is an ultra-similarity. In particular,

(i) \(G_1 \wedge G_2\) is the \(\not\sim\)-maximal graph \(G_1 \wedge G_2\) in its single-linkage equivalence class \(\forall(G_1 \wedge G_2)\) whenever both, \(G_1\) and \(G_2\) are the \(\not\sim\)-maximal graphs in their respective single-linkage equivalence classes, and

(ii) \(E_G = \binom{V}{2}\) must hold for any connected graph \(G\) that is \(\not\sim\)-maximal in its single-linkage equivalence class \(\forall(G)\) in which case also \(\omega_G^* = \omega_G\) must hold.

**Proposition 2.3.** One has \(\overline{G} \not\sim \overline{G'}\) for any graph \(G' \in W(V)\) with \(G \not\sim G'\). In other words, the map \(\omega_G^*\) associated with the \(\not\sim\)-maximal graph \(\overline{G}\) in \(\forall(G)\) referred to in Proposition 2.1 coincides with the (necessarily unique!) “smallest” map from \(\binom{V}{2}\) into \(\mathbb{R} \cup \{-\infty\}\) that is (point-wise) larger than, or equal to, \(\omega_G\) and satisfies the ultra-similarity inequality (1).

**Proposition 2.4.** The graph \(G\) is a tree if and only if it is connected and \(\not\sim\)-minimal in its single-linkage equivalence class \(\forall(G)\).

**Proposition 2.5.** If \(G\) is connected and \(T\) is an edge-weighted tree in \(W(V)\) with \(T \not\sim G\), the following three assertions are equivalent:

(i) \(T\) is a maximal spanning tree for \(G\),

(ii) for every edge \(f = \{u, v\} \in E_G\), one has \(\omega_T(f) \geq \omega_G(e)\) for every edge \(e \in E_T\) occurring in the unique path, denoted by \(P_T(u, v)\), that connects the two vertices \(u\) and \(v\) in \(T\),

(iii) \(T\) is single-linkage equivalent to \(G\).

The proof of these five propositions will be presented in Section 4 after first having collected some simple, but rather useful facts regarding the concepts introduced above.

3. Some simple useful facts

In this section, we collect, without proof, six very straightforward observations:

(O1) We have \(G = G(>t)\) for any graph \(G \in W(V)\) and any sufficiently small threshold \(t \in \mathbb{R}\). In particular, \(G\) is connected if and only if \(G(>t)\) is connected for any sufficiently small threshold \(t \in \mathbb{R}\) if and only if some or, equivalently, every graph \(G' \in \forall(G)\) is connected.

(O2) Given a graph \(G \in W(V)\) as above and a threshold \(t \in \mathbb{R}\), the graph \(G(>t)\) is totally disconnected, i.e., \(E_G(>t) = \emptyset\) or, equivalently, \(\pi_0(G(>t)) = \{v : v \in V\}\) holds, if and only if one has \(t \geq \max(\omega_G) := \max_{e \in E_G}(\omega_G(e))\).
In particular, \( \max(\omega_G) = \max(\omega_G) \) holds for any graph \( G' \in \mathbb{V}(G) \).

(03) Denoting by \( \delta(G[t]) \), for any graph \( G \in \mathbb{W}(V) \) and any \( t \in \mathbb{R} \), the set consisting of all “\( t \)-splits of \( V \) relative to \( G \)”, i.e., the set consisting of all bipartitions \( \{A, B\} \) of \( V \) into two non-empty disjoint subsets \( A, B \) for which \( \omega_G(a, b) \leq t \) holds for all \( a \in A \) and \( b \in B \) with \( \{a, b\} \in E_G \) (or, equivalently, with \( \omega_G(a, b) \leq t \) for all \( a \in A \) and \( b \in B \)), one has \( G' \subseteq G' \) for any two graphs \( G, G' \in \mathbb{W}(V) \) if and only if \( \delta(G[t]) = \delta(G'[t]) \) holds for all \( t \in \mathbb{R} \).

(04) One has \( \delta(G[t]) \subseteq \delta(G[t]) \) for any two graphs \( G, G' \in \mathbb{W}(V) \) with \( G \not\sim G' \). In particular, the single-linkage equivalence class \( \mathbb{V}(G) \) of a graph \( G \in \mathbb{W}(V) \) is a “convex subset” of \( \mathbb{W}(V) \) relative to \( \not\sim \), i.e., the “interval” \( \{G_1, G_2\} = \{1, 2\} \supset \mathbb{V}(G) \) is a subset of \( \mathbb{V}(G) \) for any two graphs \( G_1, G_2 \in \mathbb{V}(G) \).

(05) A connected graph \( G \in \mathbb{W}(V) \) is a tree if and only if \( \sup(\omega_G) \triangleq \sup(\omega_G) \) of its edges \( e \in E_G \) with \( \omega_G(e) \leq t \) coincides – for every \( t \in \mathbb{R} \) – with the number \( |\mathbb{W}(G(t))| \) if and only if this holds for every sufficiently large \( t \). In particular, given any two single-linkage equivalent trees \( T, T' \in \mathbb{W}(V) \), the number \( \sup(T(t)) \) of its edges \( e \in T \) with \( \omega_G(e) \leq t \) coincides with the correspondingly defined number for \( T' \).

(06) Given any graph \( G \in \mathbb{W}(V) \), the binary relation \( \geq \) defined on \( V \), for every \( t \in \mathbb{R} \), by \( u \geq v \iff u = v \) or \( \omega_G(u, v) \geq t \) is an equivalence relation if and only if \( \omega_G \) is an ultra-similarity. In particular, one has \( \omega_G(u_0, u_i) \geq \min(\omega_G(u_i, u_{i-1}) : i = 1, \ldots, k) \) for any sequence \( u_0, u_1, \ldots, u_k \) of distinct elements in \( V \) in this case, and \( \omega_G(u, v) \) coincides with the minimal threshold \( t \) such that \( u \) and \( v \) are not connected in \( G(t) \).

With these observations, we are now ready to prove our five main results.

4. Proofs

**Proof of Proposition 2.1.** For any subset \( \mathbb{G} \subseteq \mathbb{W}(V) \), we can construct its “supremum” \( \sup(\mathbb{G}) \) by putting \( E_\mathbb{G} \triangleq \cup_{G \in \mathbb{G}} E_G \) and \( \omega_G(e) \triangleq \sup_{G \in \mathbb{G}}(\omega_G(e)) \) for all \( e \in E_G \). Note that \( \sup(\mathbb{G}) \in \mathbb{W}(V) \) holds if and only if \( \omega_G(e) \leq C \) holds, for some constant \( C \), for all \( G \in \mathbb{G} \) and \( e \in E_G \), in which case \( G \not\sim C \) holds for some \( G' \in \mathbb{W}(V) \) for all \( G \in \mathbb{G} \) if and only if one has \( \sup(\mathbb{G}) \not\sim G' \), and that \( \delta(\sup(\mathbb{G}[t])) = \bigcap_{G \in \mathbb{G}} \delta(G[t]) \) holds in this case for all \( t \in \mathbb{R} \). Thus, if \( \mathbb{G} \subseteq \mathbb{V}(G) \) holds for some graph \( G \in \mathbb{W}(V) \), we also have \( \sup(\mathbb{G}) \in \mathbb{W}(V) \) and \( \delta(\sup(\mathbb{G}[t])) = \bigcap_{G \in \mathbb{G}} \delta(G[t]) = \bigcap_{G \in \mathbb{G}} \delta(G[t]) = \delta(G[t]) \) and, hence, \( \sup(\mathbb{G}) \in \mathbb{V}(G) \).

In particular, putting \( \mathbb{G} \triangleq \mathbb{V}(G) \), we see that \( \mathbb{V}(G) \) contains indeed a unique \( \not\sim \)-maximal graph \( \mathbb{G} \), viz. \( \mathbb{G} \triangleq \sup(\mathbb{V}(G)) \).

**Proof of Proposition 2.2.** To establish the first part of this proposition, suppose first that \( G = \mathbb{G} \) holds, i.e., \( G \) is the unique \( \not\sim \)-maximal graph in \( \mathbb{V}(G) \). Assume also, for the sake of contradiction, that there exist three distinct vertices \( u, v, w \in V \) such that \( \omega_G(u, v) < \min(\omega_G(u, w), \omega_G(w, v)) \) and, therefore, also \( (u, w), (w, v) \in E_G \), and consider the graph \( G' \) obtained from \( G \) by putting \( E_G \triangleq E_G, \omega_G(e) \triangleq \omega_G(e) \) for \( e \notin \{u, v\} \), and \( \omega_G(u, v) := \min(\omega_G(u, w), \omega_G(w, v)) \). By construction, we have \( G' \not\sim G \) and, therefore, also \( \delta(G'[t]) \subseteq \delta(G[t]) \) for all \( t \in \mathbb{R} \) in view of (04). It remains to show that also \( \delta(G[t]) \subseteq \delta(G[t]) \) holds for any \( t \in \mathbb{R} \), as this will imply \( \mathbb{V}(G') \) in contradiction to our assumption \( G = \mathbb{G} \). Indeed, \( \delta(G[t]) \subseteq \delta(G'[t]) \) clearly holds for any \( t \geq \omega_G(u, v) \); on the other hand, if \( (u, w) \in E_G \) is a t-split in \( G(t) \) for some \( t < \omega_G(u, v) \), then it is also a t-split in \( G'(t) \) because \( u \) and \( v \) are both connected to \( w \) in \( G(t) \); so \( u, v \notin \{a, b\} \) must hold for all \( a \in (A, B) \) and \( b \in B \).

Conversely, assuming that \( \omega \) is an ultra-similarity, we wish to show that \( G \) is \( \not\sim \)-maximal in \( \mathbb{V}(G) \). Otherwise, there exists a graph \( G' \in \mathbb{W}(V) \) with \( G \not\sim G' \) that is distinct from \( G \). Consider an edge \( e \in \{u, v\} \in E_G \) with \( \omega_G(e) < \omega_G^*(e) \). By (06), \( u \) and \( v \) are disconnected in \( G(t) \) for any \( t \) with \( \omega_G(e) < \omega_G^*(e) \) while they are clearly connected in \( G(t) \), a contradiction to the assumption \( \mathbb{V}(G) \).

Thus, if \( G_1, G_2 \) are two graphs in \( \mathbb{W}(V) \) that are both \( \not\sim \)-maximal graphs in their respective single-linkage equivalence classes, the corresponding maps \( \omega_1 := \omega^*_1 \) and \( \omega_2 := \omega^*_2 \) will both satisfy the ultra-similarity inequality, implying that also their (point-wise) minimum \( \omega = \omega_{G_1 \wedge G_2} \) will satisfy this inequality as

\[
\omega(u, v) = \min(\omega_1(u, v), \omega_2(u, v)) \\
\geq \min(\omega_1(u, w), \omega_2(u, w), \omega_2(u, w)) \\
= \min(\min(\omega_1(u, v), \min(\omega_1(u, w), \omega_2(u, w)))) \\
= \min(\omega(u, w), \omega(u, w))
\]

must hold for any three distinct elements \( u, v, w \in V \). So, also \( G_1 \wedge G_2 \) must be \( \not\sim \)-maximal in its single-linkage equivalence class.

Finally, note that if \( G \) is connected and the map \( \omega_G \) is an ultra-similarity, then \( E_G \) and, therefore, also \( \omega_G = \omega^*_G \) must hold in view of (06).

**Proof of Proposition 2.3.** To show that \( \mathbb{G} \not\sim \mathbb{G} \) holds for any graph \( G' \in \mathbb{W}(V) \) with \( G \not\sim G' \), it suffices to note that \( G \not\sim \mathbb{G} \wedge \mathbb{G} \not\sim \mathbb{G} \) and, therefore (cf. (04)), \( \mathbb{V}(G) \wedge \mathbb{V}(G) \) must hold in this case which in turn implies \( \mathbb{G} \wedge \mathbb{G} = \mathbb{G} \) (as there
is only one $\sim$-maximal graph in $\mathcal{W}(G)$ and $\overline{G} \sqcup \overline{G}$ must be, according to Proposition 2.2, $\sim$-maximal in its single-linkage equivalence class) and, therefore $\overline{G} = \overline{G} \sqcup \overline{G}$, as claimed. □

Proof of Proposition 2.4. If $G$ is a tree, then it is clearly connected and a $\sim$-minimal element in $\mathcal{W}(G)$ in view of (O5). Conversely, if $G$ is connected and a $\sim$-minimal element in $\mathcal{W}(G)$, we wish to show that $G$ is a tree. If this were not the case, consider any cycle $C$ contained in $G$ and let $e = \{u, v\}$ be an edge in $C$ such that its weight is minimal over all edges in $C$. Now consider the graph $G'$ obtained from $G$ by deleting $e$ (and restricting $\omega_G$ accordingly). We claim that $G' \sim G$. Clearly, we have $G' \sim G$. So, $\delta(G(t)) \subseteq \delta(G'(t))$ holds for any $t \in \mathbb{R}$. Therefore, it suffices to show that $\delta(G'(t)) \subseteq \delta(G(t))$ also holds, as this yields $G' \subseteq G$, a contradiction to the minimality of $G$. Indeed, it clearly holds for $t \geq \omega(e)$; on the other hand, if $[A, B]$ is a split in $\delta(G'(t))$ for some $t < \omega(e)$, then it is also a split in $\delta(G(t))$ because $u$ and $v$ are connected in $G'(t)$ via the remaining edges in $C$. So, $G$ must indeed be a tree. □

Proof of Proposition 2.5. We shall establish this proposition by showing that (i) $\implies$ (ii) $\implies$ (iii) $\implies$ (i) holds.

(i) $\implies$ (ii): If $\omega_G(e) > \omega_T(f)$ would hold for some edge $e = \{u, v\} \in E$ and some edge $f \in P_T(u, v)$, the tree $T'$ defined by $E_T := (E_T \setminus \{f\}) \cup \{e\}$ and $\omega_{T'} := \omega_G|_{E_{T'}}$ would have a larger weight than that of $T$. So, $\omega_G(e) \leq \omega_T(f)$ must hold, as claimed.

(ii) $\implies$ (iii): Since $T \sim G$ implies $\delta(G(t)) \subseteq \delta(T(t))$ for any $t \in \mathbb{R}$, it suffices – by (O3) – to establish conversely that, if $[A, B]$ is a split in $\delta(T(t))$, then it is also a split in $\delta(G(t))$. Yet, for any pair of vertices $\{a, b\}$ with $a \in A$ and $b \in B$, there exists (at least) one edge $\{u, v\}$ in $P_T(a, b)$ with $u \in A$ and $v \in B$ implying that $\omega_G(a, b) \leq \omega_T(u, v) \leq t$ must hold in view of (ii), as required.

(iii) $\implies$ (i): Suppose that $G' \sim T \sim G$ holds and choose a maximal spanning tree $T'$ of $G$ such that $[E_T \cap E_{T'}]$ is maximal over all maximal spanning trees of $G$. By the previous two steps, we must have $T \subseteq G' \subseteq G$ and, hence, also $T \subseteq G' \subseteq T'$. If $T'$ were not itself a maximal spanning tree of $G$, some edge $e = \{u, v\} \in E_T - E_{T'}$ would exist. Put $t := \omega_T(e)$, let $[A, B]$ be the $t$-split in $\delta(T(t))$ induced by deleting the edge $e$, and note that there must exist an edge $f = \{a, b\} \in E_T - E_{T'}$ in the unique path $P_T(u, v)$ in $T'$ with $a \in A$ and $b \in B$. In view of $[A, B] \in \delta(T(t)) = \delta(T'(t))$ and $T \sim G$, we must have $\omega_G(f) = \omega_T(f) \leq t = \omega_T(e) \leq \omega_G(e)$. So, considering the tree $T''$ defined by exchanging the edge $e \in E_T - E_{T'}$ for the edge $f \in E_{T'} - E_T$ in the maximal spanning tree $T'$, we would get yet another subtree of $G$ with total weight at least that of $T'$, but larger intersection with $T$ which is impossible. Thus, $T$ must indeed be a maximal spanning tree of $G$. □

5. Comments

In this section, we collect some comments regarding the above results. First note that the following well-known two-step strategy for single-linkage clustering – that is, for computing the hierarchy $\Pi(G) = \bigcup_{t \in \mathbb{R}} \pi_0(G(\geq t))$ of $G$-clusters – is an obvious consequence of the observations collected above: First, one determines a maximum spanning tree (or “forest”) $T$ of $G$ from the map $\omega_G$ by implementing a Fibonacci heap; then, one obtains $\Pi(T)$ and, hence, also $\Pi(G)$ by recursively removing the edges in $T$ that have smallest weight. Clearly, this algorithm has running time $O(|V|^2)$.

Next, note that the results obtained above can succinctly be presented and, thus, generalized in the context of matroid theory – a generalization that may be reported in some detail in a separate note.

And finally, note that corresponding results regarding minimal spanning trees and (subdominant) ultra-metrics [2,3] can be obtained by applying our results, given any graph $G \in \mathcal{W}(V)$ as above, to the graph $\neg G \in \mathcal{W}(V)$ defined by putting $E_{\neg G} := E_G$ and $\omega_{\neg G} := -\omega_G$, replacing the concept of $\mathcal{W}$-equivalence by the “dual” concept of $\neg\mathcal{W}$-equivalence $\sim_\neg$ defined by $G \sim_\neg G' \iff -G \sim -G' \iff \forall_{t \in \mathbb{R}} \pi_0(G(\leq t)) = \pi_0(G'(\leq t))$ where $G(\leq t)$ denotes, of course, the graph with vertex set $V$ and edge set $G(\leq t) := \{e \in E_G: \omega_G(e) \leq t\}$, and the binary relation “$\sim_\neg$” by the binary relation “$\sim$” defined, for any two graphs $G$ and $G'$ in $\mathcal{W}(V)$, by putting $G \sim_\neg G' \iff \neg G \sim \neg G' \iff E_G \subseteq E_{G'}$ and $\omega_G(e) \geq \omega_{G'}(e)$ for all $e \in E_G$. We leave the details to the reader.

Acknowledgements

The authors thank the BMBF/Germany, the Chinese Academy of Sciences, the German Max-Planck Gesellschaft, and the Institute of Advanced Study at the University of Warwick for their support.

References